Oracle Inequalities and Optimal Inference under Group Sparsity

Karim Lounici, Alexandre Tsybakov, Massimiliano Pontil, Sara A. van de Geer

Oracle Inequalities and Optimal Inference under Group Sparsity

Karim Lounici[1,2], **Massimiliano Pontil**[3], **Alexandre B. Tsybakov**[2], **Sara van de Geer**[4]

(1) Statistical Laboratory
University of Cambridge, Cambridge, UK

(2) CREST
3, Av. Pierre Larousse, 92240 Malakoff, France
{*karim.lounici,alexandre.tsybakov*}@*ensae.fr*

(3) Department of Computer Science
University College London
Gower Street, London WC1E, England, UK
E-mail: *m.pontil@cs.ucl.ac.uk*

(4) Seminar für Statistik, ETH Zürich
Rämistrasse 101, 8092 Zürich, Switzerland
geer@stat.math.ethz.ch

Abstract

We consider the problem of estimating a sparse linear regression vector β^* under a gaussian noise model, for the purpose of both prediction and model selection. We assume that prior knowledge is available on the sparsity pattern, namely the set of variables is partitioned into prescribed groups, only few of which are relevant in the estimation process. This group sparsity assumption suggests us to consider the Group Lasso method as a means to estimate β^*. We establish oracle inequalities for the prediction and ℓ_2 estimation errors of this estimator. These bounds hold under a restricted eigenvalue condition on the design matrix. Under a stronger coherence condition, we derive bounds for the estimation error for mixed $(2,p)$-norms with $1 \le p \le \infty$. When $p = \infty$, this result implies that a threshold version of the Group Lasso estimator selects the sparsity pattern of β^* with high probability. Next, we prove that the rate of convergence of our upper bounds is optimal in a minimax sense, up to a logarithmic factor, for all estimators over a class of group sparse vectors. Furthermore, we establish lower bounds for the prediction and ℓ_2 estimation errors of the usual Lasso estimator. Using this result, we demonstrate that the Group Lasso can achieve an improvement in the prediction and estimation properties as compared to the Lasso.

An important application of our results is provided by the problem of estimating multiple regression equation simultaneously or multi-task learning. In this case, our result lead to refinements of the results in [22] and allow one to establish the quantitative advantage of the Group Lasso over the usual Lasso in the multi-task setting. Finally, within the same setting, we show how our results can be extended to more general noise distributions, of which we only require the fourth moment to be finite. To obtain this extension, we establish a new maximal moment inequality, which may be of independent interest.

1 Introduction

Over the past few years there has been a great deal of attention on the problem of estimating a *sparse*[1] regression vector β^* from a set of linear measurements

$$y = X\beta^* + W. \tag{1.1}$$

Here X is a given $N \times K$ design matrix and W is a zero mean random variable modeling the presence of noise.

A main motivation behind sparse estimation comes from the observation that in several practical applications the number of variables K is much larger than the number N of observations, but the underlying model is known to be sparse, see [8, 12] and references therein. In this situation, the ordinary least squares estimator is not well-defined. A more appropriate estimation method is the ℓ_1-norm penalized least squares method, which is commonly referred to as the Lasso. The statistical properties of this estimator are now well understood, see, e.g., [4, 6, 7, 18, 21, 36] and references therein. In particular, it is possible to obtain oracle inequalities on the estimation and prediction errors, which are meaningful even in the regime $K \gg N$.

In this paper, we study the above estimation problem under additional structural conditions on the sparsity pattern of the regression vector β^*. Specifically, we assume that the set of variables can be partitioned into a number of groups, only few of which are relevant in the estimation process. In other words, not only we require that many components of the vector β^* are zero, but also that many of a priori known subsets of components are all equal to zero. This structured sparsity assumption suggests us to consider the Group Lasso method [39] as a mean to estimate β^* (see equation (2.2) below). It is based on regularization with a mixed $(2,1)$-norm, namely the sum, over the set of groups, of the square norm of the regression coefficients restricted to each of the groups. This estimator has received significant recent attention, see [3, 10, 16, 17, 19, 25, 24, 26, 28, 31] and references therein. Our principal goal is to clarify the advantage of this more stringent group sparsity assumption in the estimation process over the usual sparsity assumption. For this purpose, we shall address the issues of bounding the prediction error, the estimation error as well as estimating the sparsity pattern. The main difference from most of the previous work is that we obtain not only the upper bounds but also the corresponding lower bounds and thus establish optimal rates of estimation and prediction under group sparsity.

A main motivation for us to consider the group sparsity assumption is the practically important problem of simultaneous estimation the coefficient of multiple regression equations

$$\begin{array}{rcl} y_1 & = & X_1\beta_1^* + W_1 \\ y_2 & = & X_2\beta_2^* + W_2 \\ & \vdots & \\ y_T & = & X_T\beta_T^* + W_T. \end{array} \tag{1.2}$$

Here $X_1, \dots, X_T$ are prescribed $n \times M$ design matrices, $\beta_1^*, \dots, \beta_T^* \in \mathbb{R}^M$ are the unknown regression vectors which we wish to estimate, $y_1 \dots, y_T$ are n-dimensional vectors of observations and $W_1, \dots, W_T$ are *i.i.d.* zero mean random noise vectors. Examples in which this estimation

[1] The phrase "β^* is sparse" means that most of the components of this vector are equal to zero.

problem is relevant range from multi-task learning [2, 23, 28] and conjoint analysis [14, 20] to longitudinal data analysis [11] as well as the analysis of panel data [15, 38], among others. We briefly review these different settings in the course of the paper. In particular, multi-task learning provides a main motivation for our study. In that setting each regression equation corresponds to a different learning task; in addition to the requirement that $M \gg n$, we also allow for the number of tasks T to be much larger than n. Following [2] we assume that there are only few common important variables which are shared by the tasks. That is, we assume that the vectors $\beta_1^*, \dots, \beta_T^*$ are not only sparse but also have their sparsity patterns included in the same set of small cardinality. This group sparsity assumption induces a relationship between the responses and, as we shall see, can be used to improve estimation.

The model (1.2) can be reformulated as a single regression problem of the form (1.1) by setting $K = MT$, $N = nT$, identifying the vector β by the concatenation of the vectors $\beta_1, \dots, \beta_T$ and choosing X to be a block diagonal matrix, whose blocks are formed by the matrices $X_1, \dots, X_T$, in order. In this way the above sparsity assumption on the vectors β_t translate in a group sparsity assumption on the vector β^*, where each group is associated with one of the variables. That is, each group contains the same regression component across the different equations (1.2). Hence the results developed in this paper for the Group Lasso apply to the multi-task learning problem as a special case.

1.1 Outline of the main results

We are now ready to summarize the main contributions of this paper.

- We first establish bounds for the prediction and ℓ_2 estimation errors for the general Group Lasso setting, see Theorem 3.1. In particular, we include a "slow rate" bound, which holds under no assumption on the design matrix X. We then apply the theorem to the specific multi-task setting, leading to some refinements of the results in [22]. Specifically, we demonstrate that as the number of tasks T increases the dependence of the bound on the number of variables M disappears, provided that M grows at the rate slower than $\exp(T)$.

- We extend previous results on the selection of the sparsity pattern for the usual Lasso to the Group Lasso case, see Theorem 5.1. This analysis also allows us to establish the rates of convergence of the estimators for mixed $(2, p)$-norms with $1 \leq p \leq \infty$ (cf. Corollary5.1).

- We show that the rates of convergence in the above upper bounds for the prediction and $(2, p)$-norm estimation errors are optimal in a minimax sense (up to a logarithmic factor) for all estimators over a class of group sparse vectors β^*, see Theorem 6.1.

- We prove that the Group Lasso can achieve an improvement in the prediction and estimation properties as compared to the usual Lasso. For this purpose, we establish lower bounds for the prediction and ℓ_2 estimation errors of the Lasso estimator (cf. Theorem 7.1) and show that, in some important cases, they are greater than the corresponding upper bounds for the Group Lasso, under the same model assumptions. In particular, we clarify the advantage of the Group Lasso over the Lasso in the multi-task learning setting.

- Finally, we present an extension of the multi-task learning analysis to more general noise distributions having only bounded fourth moment, see Theorems 8.1 and 8.2; this extension is not straightforward and needs a new tool, the maximal moment inequality of Lemma 9.1, which may be of independent interest.

1.2 Previous work

Our results build upon recently developed ideas in the area of compressed sensing and sparse estimation, see, e.g., [4, 8, 12, 18] and references therein. In particular, it has been shown by different authors, under different conditions on the design matrix, that the Lasso satisfies sparsity oracle inequalities, see [4, 6, 7, 21, 18, 36, 41] and references therein. Closest to our study is the paper [4], which relies upon a Restricted Eigenvalue (RE) assumption as well as [21], which considered the problem of selection of sparsity pattern. Our techniques of proofs build upon and extend those in these papers.

Several papers analyzing statistical properties of the Group Lasso estimator appeared quite recently [3, 10, 16, 19, 25, 24, 26, 31]. Most of them are focused on the Group Lasso for additive models [16, 19, 24, 31] or generalized linear models [25]. Special choice of groups is studied in [10]. Discussion of the Group Lasso in a relatively general setting is given by Bach [3] and Nardi and Rinaldo [26]. Bach [3] assumes that the predictors (rows of matrix X) are random with a positive definite covariance matrix and proves results on consistent selection of sparsity pattern $J(\beta^*)$ when the dimension of the model (K in our case) is fixed and $N \to \infty$. Nardi and Rinaldo [26] address the issue of sparsity oracle inequalities in the spirit of [4] under the simplifying assumption that all the Gram matrices Ψ_j (see the definition below) are proportional to the identity matrix. However, the rates in their bounds are not precise enough (see comments in [22]) and they do not demonstrate advantages of the Group Lasso as compared to the usual Lasso. Obozinski et al. [28] consider the model (1.2) where all the matrices X_t are the same and all their rows are independent Gaussian random vectors with the same covariance matrix. They show that the resulting estimator achieves consistent selection of the sparsity pattern and that there may be some improvement with respect to the usual Lasso. Note that the Gaussian X_t is a rather particular example, and Obozinski et al. [28] focused on the consistent selection, rather than exploring whether there is some improvement in the prediction and estimation properties as compared to the usual Lasso. The latter issue has been addressed in our work [22] and in the parallel work of Huang and Zhang [17]. These papers considered only heuristic comparisons of the two estimators, i.e., those based on the upper bounds. Also the settings treated there did not cover the problem in whole generality. Huang and Zhang [17] considered the general Group Lasso setting but obtained only bounds for prediction and ℓ_2 estimation errors, while [22] focused only on the multi-task setting, though additionally with bounds for more general mixed $(2,p)$-norm estimation errors and consistent pattern selection properties.

1.3 Plan of the paper

This paper is organized as follows. In Section 2 we define the Group Lasso estimator and describe its application to the multi-task learning problem. In Sections 3 and 4 we study the oracle properties of this estimator in the case of Gaussian noise, presenting upper bounds on the prediction and estimation errors. In Section 5, under a stronger condition on the design matrices, we describe a simple modification of our method and show that it selects the correct sparsity pattern with an overwhelming probability. Next, in Section 6 we show that the rates of convergence in our upper bounds on prediction and $(2,p)$-norm estimation errors with $1 \le p \le \infty$ are optimal in a minimax sense, up to a logarithmic factor. In Section 7 we provide a lower bound for the Lasso estimator, which allows us to quantify the advantage of the Group Lasso over the Lasso under the group sparsity assumption. In Section 8 we discuss an extension of our results for multi-task learning to more general noise distributions. Finally, Section 9 presents a new maximal moment inequality (an extension of Nemirovski's inequality from the second to arbitrary moments), which is needed in the proofs of Section 8.

2 Method

In this section, we introduce the notation and describe the estimation method, which we analyze in the paper. We consider the linear regression model

$$y = X\beta^* + W, \tag{2.1}$$

where $\beta^* \in \mathbb{R}^K$ is the vector of regression coefficients, X is an $N \times K$ design matrix, $y \in \mathbb{R}^N$ is the response vector and $W \in \mathbb{R}^N$ is a random noise vector which will be specified later. We also denote by $x_1^\top, \dots, x_N^\top$ the rows of matrix X. Unless otherwise specified, all vectors are meant to be column vectors. Hereafter, for every positive integer ℓ, we let $\mathbb{N}_\ell$ be the set of integers from 1 and up to ℓ. Throughout the paper we assume that X is a deterministic matrix. However, it should be noted that our results extend in a standard way (as discussed, e.g., in [4], [8]) to random X satisfying the assumptions stated below with high probability.

We choose $M \le K$ and let the set $G_1, \dots, G_M$ form a prescribed partition of the index set $\mathbb{N}_K$ in M sets. That is, $\mathbb{N}_K = \cup_{j=1}^M G_j$ and, for every $j \ne j'$, $G_j \cap G_{j'} = \emptyset$. For every $j \in \mathbb{N}_M$, we let $K_j = |G_j|$ be the cardinality of G_j and denote by $\mathbf{X}_{G_j}$ the $N \times K_j$ sub-matrix of X formed by the columns indexed by G_j. We also use the notation $\Psi = X^\top X/N$ and $\Psi_j = \mathbf{X}_{G_j}^\top \mathbf{X}_{G_j}/N$ for the normalized Gram matrices of X and $\mathbf{X}_{G_j}$, respectively.

For every $\beta \in \mathbb{R}^K$ we introduce the notation $\beta^j = (\beta_k : k \in G_j)$ and, for every $1 \le p < \infty$, we define the mixed $(2,p)$-norm of β as

$$\|\beta\|_{2,p} = \left(\sum_{j=1}^M \left(\sum_{k \in G_j} \beta_k^2 \right)^{\frac{p}{2}} \right)^{\frac{1}{p}} = \left(\sum_{j=1}^M \|\beta^j\|^p \right)^{\frac{1}{p}}$$

and the $(2,\infty)$-norm of β as

$$\|\beta\|_{2,\infty} = \max_{1 \le j \le M} \|\beta^j\|,$$

where $\|\cdot\|$ is the standard Euclidean norm.

If $J \subseteq \mathbb{N}_M$ we let β_J be the vector $(\beta^j I\{j \in J\} : j \in \mathbb{N}_M)$, where $I\{\cdot\}$ denotes the indicator function. Finally we set $J(\beta) = \{j : \beta^j \ne 0,\ j \in \mathbb{N}_M\}$ and $M(\beta) = |J(\beta)|$ where $|J|$ denotes the cardinality of set $J \subset \{1, \dots, M\}$. The set $J(\beta)$ contains the indices of the relevant groups and the number $M(\beta)$ the number of such groups. Note that when $M = K$ we have $G_j = \{j\}$, $j \in \mathbb{N}_K$ and $\|\beta\|_{2,p} = \|\beta\|_p$, where $\|\beta\|_p$ is the ℓ_p norm of β.

The main assumption we make on β^* is that it is *group sparse*, which means that $M(\beta^*)$ is much smaller than M.

Our main goal is to estimate the vector β^* as well as its sparsity pattern $J(\beta^*)$ from y. To this end, we consider the Group Lasso estimator. It is defined to be a solution $\hat{\beta}$ of the optimization problem

$$\min\left\{\frac{1}{N}\|X\beta - y\|^2 + 2\sum_{j=1}^{M}\lambda_j\|\beta^j\| : \beta \in \mathbb{R}^K\right\}, \tag{2.2}$$

where $\lambda_1, \dots, \lambda_M$ are positive parameters, which we shall specify later.

In order to study the statistical properties of this estimator, it is useful to present the optimality conditions for a solution of the problem (2.2). Since the objective function in (2.2) is convex, $\hat{\beta}$ is a solution of (2.2) if and only if 0 (the K-dimensional zero vector) belongs to the subdifferential of the objective function. In turn, this condition is equivalent to the requirement that

$$-\nabla\left(\frac{1}{N}\|X\beta - y\|^2\right) \in 2\partial\left(\sum_{j=1}^{M}\lambda_j\|\hat{\beta}^j\|\right),$$

where ∂ denotes the subdifferential (see, for example, [5] for more information on convex analysis). Note that

$$\partial\left(\sum_{j=1}^{M}\lambda_j\|\beta^j\|\right) = \left\{\theta \in \mathbb{R}^K : \theta^j = \lambda_j\frac{\beta^j}{\|\beta^j\|} \text{ if } \beta^j \neq 0, \text{ and } \|\theta^j\| \leq \lambda_j \text{ if } \beta^j = 0,\ j \in \mathbb{N}_M\right\}.$$

Thus, $\hat{\beta}$ is a solution of (2.2) if and only if

$$\frac{1}{N}(X^\top(y - X\hat{\beta}))^j = \lambda_j\frac{\hat{\beta}^j}{\|\hat{\beta}^j\|}, \quad \text{if } \hat{\beta}^j \neq 0 \tag{2.3}$$

$$\frac{1}{N}\|(X^\top(y - X\hat{\beta}))^j\| \leq \lambda_j, \quad \text{if } \hat{\beta}^j = 0. \tag{2.4}$$

2.1 Simultaneous estimation of multiple regression equations and multi-task learning

As an application of the above ideas we consider the problem of estimating multiple linear regression equations simultaneously. More precisely, we consider multiple Gaussian regression models,

$$\begin{aligned} y_1 &= X_1\beta_1^* + W_1 \\ y_2 &= X_2\beta_2^* + W_2 \\ &\vdots \\ y_T &= X_T\beta_T^* + W_T, \end{aligned} \tag{2.5}$$

where, for each $t \in \mathbb{N}_T$, we let X_t be a prescribed $n \times M$ design matrix, $\beta_t^* \in \mathbb{R}^M$ the unknown vector of regression coefficients and y_t an n-dimensional vector of observations. We assume that $W_1, \dots, W_T$ are *i.i.d.* zero mean random vectors.

We study this problem under the assumption that the sparsity patterns of vectors β_t^* are for any t contained in the same set of small cardinality s. In other words, the response variable associated with each equation in (2.5) depends only on some members of a small subset of the corresponding predictor variables, which is preserved across the different equations. We consider as our estimator a solution of the optimization problem

$$\min \left\{ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \|X_t \beta_t - y_t\|^2 + 2\lambda \sum_{j=1}^{M} \left(\sum_{t=1}^{T} \beta_{tj}^2 \right)^{\frac{1}{2}} : \beta_1, \dots, \beta_T \in \mathbb{R}^M \right\} \tag{2.6}$$

with some tuning parameter $\lambda > 0$. As we have already mentioned in the introduction, this estimator is an instance of the Group Lasso estimator described above. Indeed, set $K = MT$, $N = nT$, let $\beta \in \mathbb{R}^K$ be the vector obtained by stacking the vectors $\beta_1, \dots, \beta_T$ and let y and W be the random vectors formed by stacking the vectors $y_1, \dots, y_T$ and the vectors $W_1, \dots, W_T$, respectively. We identify each row index of X with a double index $(t, i) \in \mathbb{N}_T \times \mathbb{N}_n$ and each column index with $(t, j) \in \mathbb{N}_T \times \mathbb{N}_M$. In this special case the matrix X is block diagonal and its t-th block is formed by the $n \times M$ matrix X_t corresponding to "task t". Moreover, the groups are defined as $G_j = \{(t, j) : t \in \mathbb{N}_T\}$ and the parameters λ_j in (2.2) are all set equal to a common value λ. Within this setting, we see that (2.6) is a special case of (2.2).

Finally, note that the vectors $\beta^j = (\beta_{tj} : t \in \mathbb{N}_T)^\top$ are formed by the coefficients corresponding to the j-th variable "across the tasks". The set $J(\beta) = \{j : \beta^j \neq 0,\ j \in \mathbb{N}_M\}$ contains the indices of the relevant variables present in at least one of the vectors $\beta_1, \dots, \beta_T$ and the number $M(\beta) = |J(\beta)|$ quantifies the level of group sparsity across the tasks. The structured sparsity (or group sparsity) assumption has the form $M(\beta^*) \leq s$ where s is some integer much smaller than M.

Our interest in this model with group sparsity is mainly motivated by multi-task learning. Let us briefly discuss the multi-task setting as well as other applications, in which the problem of estimating multiple regression equations arises.

Multi-task learning. In machine learning, the problem of multi-task learning has received much attention recently, see [2] and references therein. Here each regression equation corresponds to a different "learning task". In this context the tasks often correspond to binary classification, namely the response variables are binary. For instance, in image detection each task t is associated with a particular type of visual object (e.g., face, car, chair, etc.), the rows $x_{ti}^\top$ of the design matrix X_t represent an image and y_{ti} is a binary label, which, say, takes the value 1 if the image depicts the object associated with task t and the value -1 otherwise. In this setting the number of samples n is typically much smaller than the number of tasks T. A main goal of multi-task learning is to exploit possible relationships across the tasks to aid the learning process.

Conjoint analysis. In marketing research, an important problem is the analysis of datasets concerning the ratings of different products by different customers, with the purpose of improving products, see, for example, [1, 20, 14] and references therein. Here the index $t \in \mathbb{N}_T$ refers to the customers and the index $i \in \mathbb{N}_n$ refers to the different ratings provided by a customer. Products are represented by (possibly many) categorical or continuous variables (e.g., size, brand, color, price etc.). The observation y_{ti} is the rating of product x_{ti} by the t-th customer. A main goal of conjoint analysis is to find common factors which determine people's preferences to products. In this context, the variable selection method we analyze in this paper may be useful to "visualize" peoples perception of products [1].

Seemingly unrelated regressions (SUR). In econometrics, the problem of estimating the regression vectors β^*_t in (2.5) is often referred to as *seemingly unrelated regressions* (SUR) [40] (see also [34] and references therein). In this context, the index $i \in \mathbb{N}_n$ often refers to time and the equations (2.5) are equivalently represented as n systems of linear equations, indexed by time. The underlying assumption in the SUR model is that the matrices X_t are of rank M, which necessarily requires that $n \geq M$. Here we do not make such an assumption. We cover the case $n \ll M$ and show how, under a sparsity assumption, we can reliably estimate the regression vectors. The classical SUR model assumes that the noise variables are zero mean correlated Gaussian, with $\mathrm{cov}(W_s, W_t) = \sigma_{st} I_{n\times n},\ s,t \in \mathbb{N}_T$. This induces a relation between the responses that can be used to improve estimation. In our model such a relation also exists but it is described in a different way, for example, we can consider that the sparsity patterns of vectors $\beta^*_1, \ldots, \beta^*_T$ are the same.

Longitudinal and panel data. Another related context is *longitudinal data analysis* [11] as well as the analysis of *panel data* [15, 38]. Panel data refers to a dataset which contains observations of different phenomena observed over multiple instances of time (for example, election studies, political economy data, etc). The models used to analyze panel data appear to be related to the SUR model described above, but there is a large variety of model assumptions on the structure of the regression coefficients, see, for example, [15]. Up to our knowledge however, sparsity assumptions have not been been put forward for analysis within this context.

3 Sparsity oracle inequalities

Let $1 \leq s \leq M$ be an integer that gives an upper bound on the group sparsity $M(\beta^*)$ of the true regression vector β^*. We make the following assumption.

Assumption 3.1. *There exists a positive number $\kappa = \kappa(s)$ such that*

$$\min\left\{ \frac{\|X\Delta\|}{\sqrt{N}\|\Delta_J\|} \ :\ |J| \leq s, \Delta \in \mathbb{R}^K \setminus \{0\}, \sum_{j\in J^c} \lambda_j \|\Delta^j\| \leq 3 \sum_{j\in J} \lambda_j \|\Delta^j\| \right\} \geq \kappa,$$

where J^c denotes the complement of the set of indices J.

To emphasize the dependency of Assumption 3.1 on s, we will sometimes refer to it as Assumption RE(s). This is a natural extension to our setting of the Restricted Eigenvalue assumption for the usual Lasso and Dantzig selector from [4]. The ℓ_1 norms are now replaced by (weighted) mixed (2,1)-norms.

Several simple sufficient conditions for Assumption 3.1 in the Lasso case, i.e., when all the groups G_j have size 1, are given in [4]. Similar sufficient conditions can be stated in our more general setting. For example, Assumption 3.1 is immediately satisfied if $X^\top X/N$ has a positive minimal eigenvalue. More interestingly, it is enough to suppose that the matrix $X^\top X/N$ satisfies a Restricted Isometry condition as in [8] or the coherence condition (cf. Lemma A.2 below).

To state our first result we need some more notation. For every symmetric and positive semi-definite matrix A, we denote by $\mathrm{tr}(A)$, $\|A\|_{\mathrm{Fr}}$ and $|||A|||$ the trace, Frobenius and spectral norms of A, respectively. If $\rho_1, \ldots, \rho_k$ are the eigenvalues of A, we have that $\mathrm{tr}(A) = \sum_{i=1}^k \rho_i$, $\|A\|_{\mathrm{Fr}} = \sqrt{\sum_{i=1}^k \rho_i^2}$ and $|||A||| = \max_{i=1}^k \rho_i$.

Lemma 3.1. *Consider the model* (2.1)*, and let $M \geq 2$, $N \geq 1$. Assume that $W \in \mathbb{R}^N$ is a random vector with i.i.d. $\mathcal{N}(0, \sigma^2)$ gaussian components, $\sigma^2 > 0$. For every $j \in \mathbb{N}_M$, recall that $\Psi_j = \mathbf{X}_{G_j}^\top \mathbf{X}_{G_j}/N$ and choose*

$$\lambda_j \geq \frac{2\sigma}{\sqrt{N}} \sqrt{\mathrm{tr}(\Psi_j) + 2|||\Psi_j|||(2q \log M + \sqrt{K_j q \log M})}. \tag{3.1}$$

Then with probability at least $1 - 2M^{1-q}$, for any solution $\hat{\beta}$ of problem (2.2) *and all $\beta \in \mathbb{R}^K$ we have that*

$$\frac{1}{N}\|X(\hat{\beta} - \beta^*)\|^2 + \sum_{j=1}^M \lambda_j \|\hat{\beta}^j - \beta^j\| \leq \frac{1}{N}\|X(\beta - \beta^*)\|^2 + 4 \sum_{j \in J(\beta)} \lambda_j \min\left(\|\beta^j\|, \|\hat{\beta}^j - \beta^j\|\right), \tag{3.2}$$

$$\frac{1}{N}\|(X^\top X(\hat{\beta} - \beta^*))^j\| \leq \frac{3}{2}\lambda_j, \tag{3.3}$$

$$M(\hat{\beta}) \leq \frac{4\phi_{\max}}{\lambda_{\min}^2 N}\|X(\hat{\beta} - \beta^*)\|^2, \tag{3.4}$$

where $\lambda_{\min} = \min_{j=1}^M \lambda_j$ and $\phi_{\max}$ is the maximum eigenvalue of the matrix $X^\top X/N$.

Proof. For all $\beta \in \mathbb{R}^K$, we have

$$\frac{1}{N}\|X\hat{\beta} - y\|^2 + 2\sum_{j=1}^M \lambda_j \|\hat{\beta}^j\| \leq \frac{1}{N}\|X\beta - y\|^2 + 2\sum_{j=1}^M \lambda_j \|\beta^j\|,$$

which, using $y = X\beta^* + W$, is equivalent to

$$\frac{1}{N}\|X(\hat{\beta} - \beta^*)\|^2 \leq \frac{1}{N}\|X(\beta - \beta^*)\|^2 + \frac{2}{N}W^\top X(\hat{\beta} - \beta) + 2\sum_{j=1}^M \lambda_j \big(\|\beta^j\| - \|\hat{\beta}^j\|\big). \tag{3.5}$$

By the Cauchy-Schwarz inequality, we have that

$$W^\top X(\hat{\beta} - \beta) \leq \sum_{j=1}^M \|(X^\top W)^j\| \|\hat{\beta}^j - \beta^j\|.$$

For every $j \in \mathbb{N}_M$, consider the random event

$$\mathcal{A} = \bigcap_{j=1}^{M} \mathcal{A}_j, \tag{3.6}$$

where

$$\mathcal{A}_j = \left\{ \frac{1}{N} \|(X^\top W)^j\| \leq \frac{\lambda_j}{2} \right\}. \tag{3.7}$$

We note that

$$\mathbb{P}\left(\mathcal{A}_j\right) = \mathbb{P}\left(\left\{ \frac{1}{N^2} W^\top \mathbf{X}_{G_j} \mathbf{X}_{G_j}^\top W \leq \frac{\lambda_j^2}{4} \right\} \right) = \mathbb{P}\left(\left\{ \frac{\sum_{i=1}^N v_{j,i}(\xi_i^2 - 1)}{\sqrt{2}\|v_j\|} \leq x_j \right\} \right),$$

where $\xi_1, \ldots, \xi_N$ are i.i.d. standard Gaussian, $v_{j,1}, \ldots, v_{j,N}$ denote the eigenvalues of the matrix $\mathbf{X}_{G_j}\mathbf{X}_{G_j}^\top/N$, among which the positive ones are the same as those of Ψ_j, and the quantity x_j is defined as

$$x_j = \frac{\lambda_j^2 N/(4\sigma^2) - \operatorname{tr}(\Psi_j)}{\sqrt{2}\|\Psi_j\|_{\mathrm{Fr}}}.$$

We apply Lemma A.1 to upper bound the probability of the complement of the event A_j. Specifically, we choose $v = (v_{j,1}, \ldots, v_{j,N})$, $x = x_j$ and $m(v) = |||\Psi_j|||/\|\Psi_j\|_{\mathrm{Fr}}$ and conclude from Lemma A.1 that

$$\mathbb{P}\big(\mathcal{A}_j^c\big) \leq 2\exp\left(-\frac{x_j^2}{2(1 + \sqrt{2}x_j|||\Psi_j|||/\|\Psi_j\|_{\mathrm{Fr}})} \right).$$

We now choose x_j so that the right hand side of the above inequality is smaller than $2M^{-q}$. A direct computation yields that

$$x_j \geq \sqrt{2}|||\Psi_j|||/\|\Psi\|_{\mathrm{Fr}} q \log M + \sqrt{2(|||\Psi_j|||q \log M)^2 + 2q\log M},$$

which, using the subadditivity property of the square root and the inequality $\|\Psi_j\|_{\mathrm{Fr}} \leq \sqrt{K_j}|||\Psi_j|||$ gives inequality (3.1). We conclude, by a union bound, under the above condition on the parameters λ_j, that $\mathbb{P}(\mathcal{A}^c) \leq 2M^{1-q}$. Then, it follows from inequality (3.5), with probability at least $1 - 2M^{1-q}$, that

$$\begin{aligned}
\frac{1}{N}\|X(\hat\beta - \beta^*)\|^2 + \sum_{j=1}^M \lambda_j \|\hat\beta^j - \beta^j\| &\leq \frac{1}{N}\|X(\beta - \beta^*)\|^2 + 2\sum_{j=1}^M \lambda_j \big(\|\hat\beta^j - \beta^j\| + \|\beta^j\| - \|\hat\beta^j\|\big) \\
&\leq \frac{1}{N}\|X(\beta - \beta^*)\|^2 + 4\sum_{j \in J(\beta)} \lambda_j \min\Big(\|\beta^j\|, \|\hat\beta^j - \beta^j\|\Big),
\end{aligned}$$

which coincides with inequality (3.2).

To prove (3.3), we use the inequality

$$\frac{1}{N}\|(X^\top(y - X\hat\beta))^j\| \leq \lambda_j, \tag{3.8}$$

which follows from the optimality conditions (2.3) and (2.4). Moreover, using equation (2.1) and the triangle inequality, we obtain that

$$\frac{1}{N}\|(X^\top X(\hat\beta-\beta^*))^j\| \le \frac{1}{N}\|(X^\top(X\hat\beta-y))^j\| + \frac{1}{N}\|(X^\top W)^j\|.$$

The result then follows by combining the last inequality with inequality (3.8) and using the definition of the event $\mathcal{A}$.

Finally, we prove (3.4). First, observe that, on the event $\mathcal{A}$, it holds, uniformly over $j \in \mathbb{N}_M$, that

$$\frac{1}{N}\|(X^\top X(\hat\beta-\beta^*))^j\| \ge \frac{\lambda_j}{2}, \quad \text{if } \hat\beta^j \neq 0.$$

This fact follows from (2.3), (2.1) and the definition of the event $\mathcal{A}$. The following chain yields the result:

$$\begin{aligned}
M(\hat\beta) &\le \frac{4}{N^2}\sum_{j\in J(\hat\beta)} \frac{1}{\lambda_j^2}\|(X^\top X(\hat\beta-\beta^*))^j\|^2 \\
&\le \frac{4}{\lambda_{\min}^2 N^2}\sum_{j\in J(\hat\beta)} \|(X^\top X(\hat\beta-\beta^*))^j\|^2 \\
&\le \frac{4}{\lambda_{\min}^2 N^2}\|X^\top X(\hat\beta-\beta^*)\|^2 \\
&\le \frac{4\phi_{\max}}{\lambda_{\min}^2 N}\|X(\hat\beta-\beta^*)\|^2,
\end{aligned}$$

where, in the last line we have used the fact that the eigenvalues of $X^\top X/N$ are bounded from above by $\phi_{\max}$. ■

We are now ready to state the main result of this section.

Theorem 3.1. *Consider the model* (2.1) *and let $M \ge 2$, $N \ge 1$. Assume that $W \in \mathbb{R}^N$ is a random vector with i.i.d. $\mathcal{N}(0,\sigma^2)$ gaussian components, $\sigma^2 > 0$. For every $j \in \mathbb{N}_M$, define the matrix $\Psi_j = \mathbf{X}_{G_j}^\top \mathbf{X}_{G_j}/N$ and choose*

$$\lambda_j \ge \frac{2\sigma}{\sqrt{N}}\sqrt{\mathrm{tr}(\Psi_j) + 2|||\Psi_j|||(2q\log M + \sqrt{K_j q \log M})}.$$

Then with probability at least $1-2M^{1-q}$, for any solution $\hat\beta$ of problem (2.2) *we have that*

$$\frac{1}{N}\|X(\hat\beta-\beta^*)\|^2 \quad \le \quad 4\|\beta^*\|_{2,1}\max_{j=1}^{M}\lambda_j. \tag{3.9}$$

If, in addition, $M(\beta^*) \le s$ and Assumption 3.1 holds with $\kappa = \kappa(s)$, then with probability at least

$1 - 2M^{1-q}$, *for any solution* $\hat\beta$ *of problem* (2.2) *we have that*

$$\frac{1}{N}\|X(\hat\beta-\beta^*)\|^2 \;\le\; \frac{16}{\kappa^2}\sum_{j\in J(\beta^*)}\lambda_j^2, \tag{3.10}$$

$$\|\hat\beta-\beta^*\|_{2,1} \;\le\; \frac{16}{\kappa^2}\sum_{j\in J(\beta^*)}\frac{\lambda_j^2}{\lambda_{\min}}, \tag{3.11}$$

$$M(\hat\beta) \;\le\; \frac{64\phi_{\max}}{\kappa^2}\sum_{j\in J(\beta^*)}\frac{\lambda_j^2}{\lambda_{\min}^2}, \tag{3.12}$$

where $\lambda_{\min} = \min_{j=1}^M \lambda_j$ *and* $\phi_{\max}$ *is the maximum eigenvalue of the matrix* $X^\top X/N$*. If, in addition, Assumption RE(*$2s$*) holds, then with the same probability for any solution* $\hat\beta$ *of problem* (2.2) *we have that*

$$\|\hat\beta-\beta^*\| \;\le\; \frac{4\sqrt{10}}{\kappa^2(2s)}\frac{\sum_{j\in J(\beta^*)}\lambda_j^2}{\lambda_{\min}\sqrt{s}}. \tag{3.13}$$

Proof. Inequality (3.9) follows immediately from (3.2) with $\beta=\beta^*$. We now prove the remaining assertions. Let $J = J(\beta^*) = \{j : (\beta^*)^j \neq 0\}$ and let $\Delta = \hat\beta - \beta^*$. By inequality (3.2) with $\beta = \beta^*$ we have, on the event $\mathcal{A}$, that

$$\frac{1}{N}\|X\Delta\|^2 \le 4\sum_{j\in J}\lambda_j\|\Delta^j\| \le 4\sqrt{\sum_{j\in J}\lambda_j^2}\,\|\Delta_J\|. \tag{3.14}$$

Moreover by the same inequality, on the event $\mathcal{A}$, we have that $\sum_{j=1}^M \lambda_j\|\Delta^j\| \le 4\sum_{j\in J}\lambda_j\|\Delta^j\|$, which implies that $\sum_{j\in J^c}\lambda_j\|\Delta^j\| \le 3\sum_{j\in J}\lambda_j\|\Delta^j\|$. Thus, by Assumption 3.1

$$\|\Delta_J\| \le \frac{\|X\Delta\|}{\kappa\sqrt{N}}. \tag{3.15}$$

Now, (3.10) follows from (3.14) and (3.15).

Inequality (3.11) follows by noting that, by (3.2),

$$\sum_{j=1}^M \lambda_j\|\Delta^j\| \le 4\sum_{j\in J}\lambda_j\|\Delta^j\| \le 4\sqrt{\sum_{j\in J}\lambda_j^2}\|\Delta_J\| \le 4\sqrt{\sum_{j\in J}\lambda_j^2}\frac{\|X\Delta\|}{\sqrt{N}\kappa}$$

and then using (3.10) and $\sum_{j=1}^M\|\Delta^j\| \le \sum_{j=1}^M \|\Delta^j\|\lambda_j/\lambda_{\min}$.

Inequality (3.12) follows from (3.4) and (3.10).

Finally, we prove (3.13). Let J' be the set of indices in J^c corresponding to s largest values of $\lambda_j\|\Delta^j\|$. Consider the set $J_{2s} = J\cup J'$. Note that $|J_{2s}|\le 2s$. Let $j(k)$ be the index of the $k-$th largest element of the set $\{\lambda_j\|\Delta^j\| : j\in J^c\}$. Then,

$$\lambda_{j(k)}\|\Delta^{j(k)}\| \le \sum_{j\in J^c}\lambda_j\|\Delta^j\|/k.$$

This and the fact that $\sum_{j \in J^c} \lambda_j \|\Delta^j\| \leq 3 \sum_{j \in J} \lambda_j \|\Delta^j\|$ on the event $\mathcal{A}$ implies

$$\begin{aligned} \sum_{j \in J_{2s}^c} \lambda_j^2 \|\Delta^j\|^2 &\leq \sum_{k=s+1}^{\infty} \frac{\left(\sum_{\ell \in J^c} \lambda_\ell \|\Delta^\ell\|\right)^2}{k^2} \\ &\leq \frac{\left(\sum_{\ell \in J^c} \lambda_\ell \|\Delta^\ell\|\right)^2}{s} \leq \frac{9\left(\sum_{\ell \in J} \lambda_\ell \|\Delta^\ell\|\right)^2}{s} \\ &\leq \frac{9(\sum_{j \in J} \lambda_j^2)\|\Delta_J\|^2}{s} \leq \frac{9(\sum_{j \in J} \lambda_j^2)\|\Delta_{J_{2s}}\|^2}{s}. \end{aligned}$$

Therefore, it follows that

$$\lambda_{\min}^2 \|\Delta_{J_{2s}^c}\|^2 \leq \frac{9}{s} \sum_{j \in J} \lambda_j^2 \|\Delta_{J_{2s}}\|^2$$

and, in turn, that

$$\|\Delta\|^2 \leq \frac{10}{s} \sum_{j \in J} \frac{{\lambda_j}^2}{\lambda_{\min}^2} \|\Delta_{J_{2s}}\|^2. \tag{3.16}$$

Next note from (3.14) that

$$\frac{1}{N}\|X\Delta\|^2 \leq 4\sqrt{\sum_{j \in J} \lambda_j^2}\|\Delta_{J_{2s}}\|. \tag{3.17}$$

In addition, $\sum_{j \in J^c} \lambda_j \|\Delta^j\| \leq 3 \sum_{j \in J} \lambda_j \|\Delta^j\|$ easily implies that

$$\sum_{j \in J_{2s}^c} \lambda_j \|\Delta^j\| \leq 3 \sum_{j \in J_{2s}} \lambda_j \|\Delta^j\|.$$

Combining Assumption RE($2s$) with (3.17) we have, on the event $\mathcal{A}$, that

$$\|\Delta_{J_{2s}}\| \leq \frac{4\sqrt{\sum_{j \in J} \lambda_j^2}}{\kappa^2(2s)}.$$

This inequality and (3.16) yield (3.13). ∎

The oracle inequality (3.10) of Theorem 3.1 can be generalized to include the bias term as follows.

Theorem 3.2. *Let the assumptions of Lemma 3.1 be satisfied and let Assumption 3.1 holds with $\kappa = \kappa(s)$ and with factor 3 replaced by 7. Then with probability at least $1 - 2M^{1-q}$, for any solution $\hat{\beta}$ of problem* (2.2) *we have*

$$\frac{1}{N}\|X(\hat{\beta} - \beta^*)\|^2 \leq \min\left\{ \frac{96}{\kappa^2} \sum_{j \in J(\beta)} \lambda_j^2 + \frac{2}{N}\|X(\beta - \beta^*)\|^2 : \beta \in \mathbb{R}^K, M(\beta) \leq s \right\}.$$

This result is of interest when β^* is only assumed to approximately sparse, that is when there exists a set of indices J_0 with cardinality smaller than s such that $\|(\beta^*)_{J_0^c}\|^2$ is small.

Proof. Let β be arbitrary. Set $\Delta = \hat{\beta} - \beta$. By inequality (3.2), we have, on the event $\mathcal{A}$ that

$$\frac{1}{N}\|X(\hat{\beta}-\beta^*)\|^2 + \sum_{j=1}^{M}\lambda_j\|\Delta^j\| \leq \frac{1}{N}\|X(\beta-\beta^*)\|^2 + 4\sum_{j\in J(\beta)}\lambda_j\|\Delta^j\|.$$

Let $y > 0$ be arbitrary. We consider two cases:

case i) $4\sum_{j\in J(\beta)}\lambda_j\|\Delta^j\| \geq \frac{1}{N}\|X(\beta-\beta^*)\|^2$

case ii) $4\sum_{j\in J(\beta)}\lambda_j\|\Delta^j\| < \frac{1}{N}\|X(\beta-\beta^*)\|^2$

In case i), we have

$$\frac{1}{N}\|X(\hat{\beta}-\beta^*)\|^2 + \sum_{j=1}^{M}\lambda_j\|\Delta^j\| \leq 8\sum_{j\in J(\beta)}\lambda_j\|\Delta^j\|.$$

This implies

$$\sum_{j\in J(\beta)^c}\lambda_j\|\Delta^j\| < 7\sum_{j\in J(\beta)}\lambda_j\|\Delta^j\|.$$

Thus, by Assumption 3.1 (with factor 3 replaced by 7), we have

$$\|\Delta_{J(\beta)}\| \leq \frac{\|X\Delta\|}{\kappa\sqrt{N}}.$$

We obtain

$$\begin{aligned}
\frac{1}{N}\|X(\hat{\beta}-\beta^*)\|^2 + \sum_{j=1}^{M}\lambda_j\|\Delta^j\| &\leq \frac{8}{\kappa}\sqrt{\sum_{j\in J(\beta)}\lambda_j^2}\,\frac{\|X\Delta\|}{\sqrt{N}} \\
&\leq \frac{8}{\kappa}\sqrt{\sum_{j\in J(\beta)}\lambda_j^2}\left[\frac{\|X(\hat{\beta}-\beta^*)\|}{\sqrt{N}} + \frac{\|X(\beta-\beta^*)\|}{\sqrt{N}}\right] \\
&\leq \frac{1}{2}\frac{\|X(\hat{\beta}-\beta^*)\|^2}{N} + \frac{32}{\kappa^2}\sum_{j\in J(\beta)}\lambda_j^2 \\
&\qquad + \frac{\|X(\beta-\beta^*)\|^2}{N} + \frac{16}{\kappa^2}\sum_{j\in J(\beta)}\lambda_j^2.
\end{aligned}$$

Hence

$$\frac{1}{N}\|X(\hat{\beta}-\beta^*)\|^2 + 2\sum_{j=1}^{M}\lambda_j\|\Delta^j\| \leq \frac{96}{\kappa^2}\sum_{j\in J(\beta)}\lambda_j^2 + \frac{2}{N}\|X(\beta-\beta^*)\|^2.$$

Case ii) gives

$$\frac{1}{N}\|X(\hat{\beta}-\beta^*)\|^2 + \sum_{j=1}^{M} \lambda_j \|\Delta^j\| < \frac{2}{N}\|X(\beta-\beta^*)\|^2.$$

Hence

$$\frac{1}{N}\|X(\hat{\beta}-\beta^*)\|^2 \le \min_{\beta}\left[\frac{96}{\kappa^2}\sum_{j\in J(\beta)} \lambda_j^2 + \frac{2}{N}\|X(\beta-\beta^*)\|^2\right].$$

■

We end this section by a remark about the Group Lasso estimator with overlapping groups, i.e., when $\mathbb{N}_K = \cup_{j=1}^{M} G_j$ but $G_j \cap G_{j'} \neq \emptyset$ for some $j, j' \in \mathbb{N}_M$, $j \neq j'$. We refer to [42] for motivation and discussion featuring the statistical relevance of group sparsity with overlapping groups. Inspection of the proofs of Lemma 3.1 and Theorem 3.1 immediately yields the following conclusion.

Remark 3.1. *Inequalities* (3.2) *and* (3.3) *in Lemma 3.1 and inequalities* (3.10)–(3.12) *in Theorem 3.1 remain correct in the more general case of overlapping groups* $G_1, \dots, G_M$.

4 Sparsity oracle inequalities for multi-task learning

We now apply the above results to the multi-task learning problem described in Section 2.1. In this setting, $K = MT$ and $N = nT$, where T is the number of tasks, n is the sample size for each task and M is the nominal dimension of unknown regression parameters for each task. Also, for every $j \in \mathbb{N}_M$, $K_j = T$ and $\Psi_j = (1/T)I_{T\times T}$, where $I_{T\times T}$ is the $T \times T$ identity matrix. This fact is a consequence of the block diagonal structure of the design matrix X and the assumption that the variables are normalized to one, namely all the diagonal elements of the matrix $(1/n)X^\top X$ are equal to one. It follows that $\mathrm{tr}(\Psi_j) = 1$ and $|||\Psi_j||| = 1/T$. The regularization parameters λ_j are all equal to the same value λ, cf. (2.6). Therefore, (3.1) takes the form

$$\lambda \ge \frac{2\sigma}{\sqrt{nT}}\sqrt{1 + \frac{2}{T}\left(2q\log M + \sqrt{Tq\log M}\right)}. \tag{4.1}$$

In particular, Lemma 3.1 and Theorem 3.1 are valid for

$$\lambda \ge \frac{2\sqrt{2}\sigma}{\sqrt{nT}}\sqrt{1 + \frac{5q}{2}\frac{\log M}{T}}$$

since the right-hand side of this inequality is greater than that of (4.1).

For the convenience of the reader we state the Restricted Eigenvalue assumption for the multi-task case [22].

Assumption 4.1. *There exists a positive number* $\kappa_{\mathrm{MT}} = \kappa_{\mathrm{MT}}(s)$ *such that*

$$\min\left\{\frac{\|X\Delta\|}{\sqrt{n}\|\Delta_J\|} \;:\; |J| \le s, \Delta \in \mathbb{R}^{MT}\setminus\{0\}, \|\Delta_{J^c}\|_{2,1} \le 3\|\Delta_J\|_{2,1}\right\} \ge \kappa_{\mathrm{MT}},$$

where J^c *denotes the complement of the set of indices* J.

We note that parameters $\kappa, \phi_{\max}$ defined in Section 3 correspond to $\kappa_{\rm MT}/\sqrt{T}$ and $\phi_{\rm MT}/T$ respectively, where $\phi_{\rm MT}$ is the largest eigenvalue of the matrix $X^\top X/n$.

Using the above observations we obtain the following corollary of Theorem 3.1.

Corollary 4.1. *Consider the multi-task model* (2.5) *for $M \geq 2$ and $T, n \geq 1$. Assume that $W \in \mathbb{R}^N$ is a random vector with i.i.d. $\mathcal{N}(0, \sigma^2)$ gaussian components, $\sigma^2 > 0$, and all diagonal elements of the matrix $X^\top X/n$ are equal to* 1. *Set*

$$\lambda = \frac{2\sqrt{2}\sigma}{\sqrt{nT}}\left(1 + \frac{A\log M}{T}\right)^{1/2},$$

where $A > 5/2$. Then with probability at least $1 - 2M^{1-2A/5}$, for any solution $\hat\beta$ of problem (2.6) *we have that*

$$\frac{1}{nT}\|X(\hat\beta - \beta^*)\|^2 \leq \ \frac{8\sqrt{2}\sigma}{\sqrt{nT}}\left(1 + \frac{A\log M}{T}\right)^{1/2} \|\beta^*\|_{2,1}\,. \tag{4.2}$$

Moreover, if in addition it holds that $M(\beta^*) \leq s$ and Assumption 4.1 holds with $\kappa_{\rm MT} = \kappa_{\rm MT}(s)$, then with probability at least $1 - 2M^{1-2A/5}$, for any solution $\hat\beta$ of problem (2.6) *we have that*

$$\frac{1}{nT}\|X(\hat\beta - \beta^*)\|^2 \leq \frac{128\sigma^2}{\kappa_{\rm MT}^2}\frac{s}{n}\left(1 + \frac{A\log M}{T}\right) \tag{4.3}$$

$$\frac{1}{\sqrt{T}}\|\hat\beta - \beta^*\|_{2,1} \leq \frac{32\sqrt{2}\sigma}{\kappa_{\rm MT}^2}\frac{s}{\sqrt{n}}\left(1 + \frac{A\log M}{T}\right)^{1/2} \tag{4.4}$$

$$M(\hat\beta) \leq \frac{64\phi_{\rm MT}}{\kappa_{\rm MT}^2}s, \tag{4.5}$$

where $\phi_{\rm MT}$ is the largest eigenvalue of the matrix $X^\top X/n$.

Finally, if in addition $\kappa_{\rm MT}(2s) > 0$, then with the same probability for any solution $\hat\beta$ of problem (2.6) *we have that*

$$\frac{1}{\sqrt{T}}\|\hat\beta - \beta^*\| \leq \frac{16\sqrt{5}\sigma}{\kappa_{\rm MT}^2(2s)}\sqrt{\frac{s}{n}}\left(1 + \frac{A\log M}{T}\right)^{1/2}. \tag{4.6}$$

Note that the values T and $\sqrt{T}$ in the denominators of the left-hand sides of inequalities (4.3), (4.4), and (4.6) appear quite naturally. For instance, the norm $\|\hat\beta - \beta^*\|_{2,1}$ in (4.4) is a sum of M terms each of which is a Euclidean norm of a vector in $\mathbb{R}^T$, and thus it is of the order $\sqrt{T}$ if all the components are equal. Therefore, (4.4) can be interpreted as a correctly normalized "error per coefficient" bound.

Corollary 4.1 is valid for any fixed n, M, T; the approach is non-asymptotic. Some relations between these parameters are relevant in the particular applications and various asymptotics can be derived as special cases. For example, in multi-task learning it is natural to assume that $T \geq n$, and the motivation for our approach is the strongest if also $M \gg n$. The bounds of Corollary 4.1 are meaningful if the sparsity index s is small as compared to the sample size n and the logarithm of the dimension $\log M$ is not too large as compared to T.

More interestingly, the dependency on the dimension M in the bounds is negligible if the number of tasks T is larger than $\log M$. In this regime, no relation between the sample size n and the dimension M is required. This is quite in contrast to the standard results on sparse recovery where the condition

$$\log(\text{dimension}) \quad \ll \quad \text{sample size}$$

is considered as *sine qua non* constraint. For example, Corollary 4.1 gives meaningful bounds if $M = \exp(n^\gamma)$ for arbitrarily large $\gamma > 0$, provided that $T > n^\gamma$.

Finally, note that Corollary 4.1 is in the same spirit as a result that we obtained in [22] but there are two important differences. First, in [22] we considered larger values of λ, namely with $\left(1 + \frac{A \log M}{\sqrt{T}}\right)^{1/2}$ in place of $\left(1 + \frac{A \log M}{T}\right)^{1/2}$, and we obtained a result with higher probability. We switch here to the smaller λ since it leads to minimax rate optimality, cf. lower bounds below. The second difference is that we include now the "slow rate" result (4.2), which guarantees convergence of the prediction loss *with no restriction on the matrix* $X^\top X$, provided that the norm $(2,1)$-norm of β^* is bounded. For example, if the absolute values of all components of β^* do not exceed some constant $\beta_{\max}$, then $\|\beta^*\|_{2,1} \le \beta_{\max} s \sqrt{T}$ and the bound (4.2) is of the order $\frac{s}{\sqrt{n}} \left(1 + \frac{A \log M}{T}\right)^{1/2}$.

5 Coordinate-wise estimation and selection of sparsity pattern

In this section we show how from any solution of (2.2), we can estimate the correct sparsity pattern $J(\beta^*)$ with high probability. We also establish bounds for estimation of β^* in all $(2,p)$ norms with $1 \le p \le \infty$ under a stronger condition than Assumption 3.1.

Recall that we use the notation $\Psi = \frac{1}{N} X^\top X$ for the Gram matrix of the design. We introduce some additional notation which will be used throughout this section. For any j, j' in $\mathbb{N}_M$ we define the matrix $\Psi[j,j'] = \frac{1}{N} \mathbf{X}_{G_j}^\top \mathbf{X}_{G_{j'}}$ (note that $\Psi[j,j] = \Psi_j$ for any j). We denote by $\Psi[j,j']_{t,t'}$, where $t \in \mathbb{N}_{K_j}, t' \in \mathbb{N}_{K_{j'}}$, the (t,t')-th element of matrix $\Psi[j,j']$. For any $\Delta \in \mathbb{R}^K$ and $j \in \mathbb{N}_M$ we set $\Delta^j = (\Delta_t : t \in \mathbb{N}_{K_j})$.

In this section, we assume that the following condition holds true.

Assumption 5.1. *There exist some integer $s \ge 1$ and some constant $\alpha > 0$ such that:*

1. *For any $j \in \mathbb{N}_M$ and $t \in \mathbb{N}_{K_j}$ it holds that $(\Psi[j,j])_{t,t} = \phi$ and*

$$\max_{1 \le t,t' \le K_j, t \ne t'} |(\Psi[j,j])_{t,t'}| \le \frac{\lambda_{\min} \phi}{14 \alpha \lambda_{\max} s} \frac{1}{\sqrt{K_j K_{j'}}}.$$

2. *For any $j \ne j' \in \mathbb{N}_M$ it holds that*

$$\max_{1 \le t \le \min(K_j, K_{j'})} |(\Psi[j,j'])_{t,t}| \le \frac{\lambda_{\min} \phi}{14 \alpha \lambda_{\max} s}$$

and

$$\max_{1 \le t \le K_j, 1 \le t' \le K_{j'}, t \ne t'} |(\Psi[j,j'])_{t,t'}| \le \frac{\lambda_{\min} \phi}{14 \alpha \lambda_{\max} s} \frac{1}{\sqrt{K_j K_{j'}}}.$$

This assumption is an extension to the general Group Lasso setting of the coherence condition of [22] introduced in the particular multi-task setting. Indeed, in the multi-task case $K_j \equiv T$, $\lambda_{\min} = \lambda_{\max}$, and for any $j \in \mathbb{N}_M$ the matrix $\mathbf{X}_{G_j}$ is block diagonal with the t-th block of size $n \times 1$ formed by the j-th column of the matrix X_t (recall the notation in Section 2.1) and $\phi = 1/T$. It follows that $(\Psi[j,j'])_{t,t'} = 0$ for any $j, j' \in \mathbb{N}_M$ and $t \neq t' \in \mathbb{N}_T$. Then Assumption 5.1 reduces to the following: $\max_{1\leq t \leq T} |(\Psi[j,j'])_{t,t}| \leq \frac{1}{14\alpha s T}$ whenever $j \neq j'$ and $(\Psi[j,j])_{t,t} = \frac{1}{T}$. Thus, we see that for the multi-task model Assumption 5.1 takes the form of the usual coherence assumption for each of the T separate regression problems. We also note that, the coherence assumption in [22] was formulated with the numerical constant 7 instead of 14. The larger constant here is due to the fact that we consider the general model with not necessarily block diagonal design matrix, in contrast to the multi-task setting of [22].

Lemma A.2, which is presented in the appendix, establishes that Assumption 5.1 implies Assumption 3.1. Note also that, by an argument as in [21], it is not hard to show that under Assumption 5.1 any group s-sparse vector β^* satisfying (2.1) is unique.

Theorem 3.1 provides bounds for compound measures of risk, that is, depending simultaneously on all the vectors β^j. An important question is to evaluate the performance of estimators for each of the components β^j separately. The next theorem provides a bound of this type and, as a consequence, a result on the selection of sparsity pattern.

Theorem 5.1. *Let the assumptions of Theorem 3.1 be satisfied and let Assumption 5.1 hold with the same s. Set*

$$c = \left(\frac{3}{2} + \frac{16}{7(\alpha - 1)}\right). \tag{5.1}$$

Then with probability at least $1 - 2M^{1-q}$, for any solution $\hat{\beta}$ of problem (2.2) we have that

$$\|\hat{\beta} - \beta^*\|_{2,\infty} \leq \frac{c}{\phi}\lambda_{\max}. \tag{5.2}$$

If, in addition,

$$\min_{j \in J(\beta^*)} \|(\beta^*)^j\| > \frac{2c}{\phi}\lambda_{\max}, \tag{5.3}$$

then with the same probability for any solution $\hat{\beta}$ of problem (2.2) the set of indices

$$\hat{J} = \left\{ j : \|\hat{\beta}^j\| > \frac{c}{\phi}\lambda_{\max} \right\} \tag{5.4}$$

estimates correctly the sparsity pattern $J(\beta^*)$, that is,

$$\hat{J} = J(\beta^*).$$

Proof. Set $K_\infty = \max_{1 \leq j \leq M} K_j$. We define first for any $j, j' \in \mathbb{N}_M$ the $K_\infty \times K_\infty$ matrix $\tilde{\Psi}[j,j']$ as follows. If $j \neq j'$ we have $(\tilde{\Psi}[j,j'])_{t \in \mathbb{N}_{K_j}, t' \in \mathbb{N}_{K_{j'}}} = \Psi[j,j']$ and $(\tilde{\Psi}[j,j'])_{t,t'} = 0$ if $t > K_j$ or if $t' > K_{j'}$. If $j = j'$ we have $(\tilde{\Psi}[j,j])_{t,t' \in \mathbb{N}_{K_j}} = \Psi[j,j] - \phi I_{K_j \times K_j}$ and $(\tilde{\Psi}[j,j])_{t,t'} = 0$ if $t > K_j$ or if $t' > K_j$. Similarly, for any $\Delta \in \mathbb{R}^K$ and any $j \in \mathbb{N}_M$ we set $\tilde{\Delta}^j \in \mathbb{R}^{K_\infty}$ such that $(\tilde{\Delta}^j_t)_{t \in \mathbb{N}_{K_j}} = \Delta^j$ and $\tilde{\Delta}^j_t = 0$ for any $t > K_j$.

Set $\Delta = \hat{\beta} - \beta^*$. We have

$$\phi\|\Delta\|_{2,\infty} \leq \|\Psi\Delta\|_{2,\infty} + \|(\Psi - \phi I_{K\times K})\Delta\|_{2,\infty}. \tag{5.5}$$

Using Cauchy-Schwarz's inequality we obtain

$$\begin{aligned}
\|(\Psi - \phi I_{K\times K})\Delta\|_{2,\infty} &= \max_{1\leq j\leq M}\left[\sum_{t=1}^{K_j}\left(\sum_{j'=1}^{M}\sum_{t'=1}^{K_{j'}}\left(\tilde{\Psi}[j,j']\right)_{t,t'}\tilde{\Delta}_{t'}^{j'}\right)^2\right]^{1/2} \\
&\leq \max_{1\leq j\leq M}\left[\sum_{t=1}^{K_j}\left(\sum_{j'=1}^{M}\left(\tilde{\Psi}[j,j']\right)_{t,t}\tilde{\Delta}_{t}^{j'}\right)^2\right]^{1/2} \\
&\quad + \max_{1\leq j\leq M}\left[\sum_{t=1}^{K_j}\left(\sum_{j'=1}^{M}\sum_{t'=1,t'\neq t}^{K_{j'}}\left(\tilde{\Psi}[j,j']\right)_{t,t'}\tilde{\Delta}_{t'}^{j'}\right)^2\right]^{1/2}.
\end{aligned} \tag{5.6}$$

We now treat the first term on the right-hand side of (5.6). We have, using Assumption 5.1 and Minkowski's inequality for the Euclidean norm in $\mathbb{R}^{K_j}$, that

$$\begin{aligned}
\max_{1\leq j\leq M}\left[\sum_{t=1}^{K_j}\left(\sum_{j'=1}^{M}\left(\tilde{\Psi}[j,j']\right)_{t,t}\tilde{\Delta}_{t}^{j'}\right)^2\right]^{1/2} &\leq \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\left[\sum_{t=1}^{K_j}\left(\sum_{j'=1}^{M}|\tilde{\Delta}_{t}^{j'}|\right)^2\right]^{1/2} \\
&\leq \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\|\tilde{\Delta}\|_{2,1} \\
&\leq \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\|\Delta\|_{2,1},
\end{aligned}$$

since $\|\tilde{\Delta}\|_{2,1} \leq \|\Delta\|_{2,1}$ by definition of $\tilde{\Delta}$. Next we treat the second term in the right-hand side of (5.6). Cauchy-Schwarz's inequality gives

$$\begin{aligned}
&\max_{1\leq j\leq M}\left[\sum_{t=1}^{K_j}\left(\sum_{j'=1}^{M}\sum_{t'=1,t'\neq t}^{K_{j'}}\left(\tilde{\Psi}[j,j']\right)_{t,t'}\tilde{\Delta}_{t'}^{j'}\right)^2\right]^{1/2} \\
&\qquad \leq \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\max_{1\leq j\leq M}\left[\frac{1}{K_j}\sum_{t=1}^{K_j}\left(\sum_{j'=1}^{M}\sum_{t'=1}^{K_{j'}}\frac{|\tilde{\Delta}_{t'}^{j'}|}{\sqrt{K_{j'}}}\right)^2\right]^{1/2} \\
&\qquad \leq \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\sum_{j'=1}^{M}\sum_{t'=1}^{K_{j'}}\frac{|\tilde{\Delta}_{t'}^{j'}|}{\sqrt{K_{j'}}} \\
&\qquad \leq \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\|\tilde{\Delta}\|_{2,1} \leq \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\|\Delta\|_{2,1}.
\end{aligned}$$

Combining the four above displays we get

$$\|\Delta\|_{2,\infty} \leq \frac{1}{\phi}\|\Psi\Delta\|_{2,\infty} + \frac{2\lambda_{\min}}{14\alpha\lambda_{\max}s}\|\Delta\|_{2,1}.$$

Thus, by inequalities (3.3) and (3.11), with probability at least $1 - 2M^{1-q}$, it holds that

$$\|\Delta\|_{2,\infty} \le \left(\frac{3}{2\phi} + \frac{16}{7\alpha\kappa^2}\right)\lambda_{\max}.$$

By Lemma A.2, $\alpha\kappa^2 = (\alpha - 1)\phi$, which yields the first result of the theorem. The second result follows from the first one in an obvious way. ■

Assumption of type (5.3) is inevitable in the context of selection of sparsity pattern. It says that the vectors $(\beta^*)^j$ cannot be arbitrarily close to 0 for j in the pattern. Their norms should be at least somewhat larger than the noise level.

Theorems 3.1 and 5.1 imply the following corollary.

Corollary 5.1. *Let the assumptions of Theorem 3.1 be satisfied and let Assumption 5.1 hold with the same s. Then with probability at least $1 - 2M^{1-q}$, for any solution $\hat\beta$ of problem (2.2) and any $1 \le p < \infty$ we have that*

$$\|\hat\beta - \beta^*\|_{2,p} \le \frac{c_1}{\phi}\lambda_{\max}\left(\sum_{j\in J(\beta^*)}\frac{\lambda_j^2}{\lambda_{\min}\lambda_{\max}}\right)^{\frac{1}{p}}, \tag{5.7}$$

where

$$c_1 = \left(\frac{16\alpha}{\alpha - 1}\right)^{1/p}\left(\frac{3}{2} + \frac{16}{7(\alpha - 1)}\right)^{1-\frac{1}{p}}. \tag{5.8}$$

If, in addition, (5.3) holds, then with the same probability for any solution $\hat\beta$ of problem (2.2) and any $1 \le p < \infty$ we have that

$$\|\hat\beta - \beta^*\|_{2,p} \le \frac{c_1}{\phi}\lambda_{\max}\left(\sum_{j\in\hat J}\frac{\lambda_j^2}{\lambda_{\min}\lambda_{\max}}\right)^{\frac{1}{p}}, \tag{5.9}$$

where $\hat J$ is defined in (5.4).

Proof. Set $\Delta = \hat\beta - \beta$. For any $p \ge 1$ we use the norm interpolation inequality

$$\|\Delta\|_{2,p} \le \|\Delta\|_{2,1}^{\frac{1}{p}}\|\Delta\|_{2,\infty}^{1-\frac{1}{p}}.$$

Combining inequalities (3.11) and (5.2) with $\kappa = \sqrt{(1 - 1/\alpha)\phi}$ (cf. Lemma A.2) and the last inequality yields (5.7). Inequality (5.9) is then straightforward in view of Theorem 5.1. ■

Note that we introduce inequalities (5.2) and (5.9) valid with probability close to 1 because their right-hand sides are data driven, and so they can be used as confidence bands for the unknown parameter β^* in mixed (2,p)-norms.

We finally derive a corollary of Theorem 5.1 for the multi-task setting, which is straightforward in view of the above results.

Corollary 5.2. *Consider the multi-task model* (2.5) *for $M \geq 2$ and $T, n \geq 1$. Let the assumptions of Theorem 5.1 be satisfied and set*

$$\lambda = \frac{2\sqrt{2}\sigma}{\sqrt{nT}} \left(1 + \frac{A \log M}{T}\right)^{1/2},$$

where $A > 5/2$. Then with probability at least $1 - 2M^{1-2A/5}$, for any solution $\hat{\beta}$ of problem (2.6) and any $1 \leq p \leq \infty$ we have

$$\frac{1}{\sqrt{T}} \|\hat{\beta} - \beta^*\|_{2,p} \leq \frac{2\sqrt{2} c_1 \sigma s^{1/p}}{\sqrt{n}} \left(1 + \frac{A \log M}{T}\right)^{1/2}, \tag{5.10}$$

where c_1 is the constant defined in (5.8) and we set $x^{1/\infty} = 1$ for any $x > 0$. If, in addition,

$$\min_{j \in J(\beta^*)} \frac{1}{\sqrt{T}} \|(\beta^*)^j\| > \frac{4\sqrt{2} c\sigma}{\sqrt{n}} \left(1 + \frac{A \log M}{T}\right)^{1/2}, \tag{5.11}$$

then with the same probability for any solution $\hat{\beta}$ of problem (2.6) the set of indices

$$\hat{J} = \left\{ j : \frac{1}{\sqrt{T}} \|\hat{\beta}^j\| > \frac{2\sqrt{2} c\sigma}{\sqrt{n}} \left(1 + \frac{A \log M}{T}\right)^{1/2} \right\} \tag{5.12}$$

estimates correctly the sparsity pattern $J(\beta^*)$, that is,

$$\hat{J} = J(\beta^*).$$

6 Minimax lower bounds for arbitrary estimators

In this section we consider again the multi-task model as in Sections 2.1 and 4. We will show that the rate of convergence obtained in Corollary 4.1 is optimal in a minimax sense (up to a logarithmic factor) for all estimators over a class of group sparse vectors. This will be done under the following mild condition on matrix X.

Assumption 6.1. *There exist positive constants κ_1 and κ_2 such that for any vector $\Delta \in \mathbb{R}^{MT} \setminus \{0\}$ with $M(\Delta) \leq 2s$ we have*

$$(a) \quad \frac{\|X\Delta\|^2}{n\|\Delta\|^2} \geq \kappa_1^2, \qquad\qquad (b) \quad \frac{\|X\Delta\|^2}{n\|\Delta\|^2} \leq \kappa_2^2.$$

Note that part (b) of Assumption 6.1 is automatically satisfied with $\kappa_2^2 = \phi_{\mathrm{MT}}$ where ϕ_{MT} is the spectral norm of matrix $X^\top X/n$. The reason for introducing this assumption is that the $2s$-restricted maximal eigenvalue κ_2^2 can be much smaller than the spectral norm of $X^\top X/n$, which would result in a sharper lower bound, see Theorem 6.1 below.

In what follows we fix $T \geq 1$, $M \geq 2$, $s \leq M/2$ and denote by $GS(s, M, T)$ the set of vectors $\beta \in \mathbb{R}^{MT}$ such that $M(\beta) \leq s$. Let $\ell : \mathbb{R}^+ \to \mathbb{R}^+$ be a nondecreasing function such that $\ell(0) = 0$ and $\ell \not\equiv 0$.

Theorem 6.1. *Consider the multi-task model* (2.5) *for* $M \geq 2$ *and* $T, n \geq 1$*. Assume that* $W \in \mathbb{R}^N$ *is a random vector with i.i.d.* $\mathcal{N}(0,\sigma^2)$ *gaussian components,* $\sigma^2 > 0$*. Suppose that* $s \leq M/2$ *and let part (b) of Assumption 6.1 be satisfied. Define*

$$\psi_{n,p} = \frac{\sigma}{\kappa_2}\frac{s^{1/p}}{\sqrt{n}}\left(1 + \frac{\log(eM/s)}{T}\right)^{1/2}, \qquad 1 \leq p \leq \infty,$$

where we set $s^{1/\infty} = 1$*. Then there exist positive constants* $\bar{b}, \bar{c}$ *depending only on* $\ell(\cdot)$ *and* p *such that*

$$\inf_{\tau} \sup_{\beta^* \in GS(s,M,T)} \mathbb{E}\ell\left(\bar{b}\psi_{n,p}^{-1}\frac{1}{\sqrt{T}}\|\tau - \beta^*\|_{2,p}\right) \geq \bar{c}, \tag{6.1}$$

where $\inf_\tau$ *denotes the infimum over all estimators* τ *of* β^**. If, in addition, part (a) of Assumption 6.1 is satisfied, then there exist positive constants* $\bar{b}, \bar{c}$ *depending only on* $\ell(\cdot)$ *such that*

$$\inf_{\tau} \sup_{\beta^* \in GS(s,M,T)} \mathbb{E}\ell\left(\bar{b}\psi_{n,2}^{-1}\frac{1}{\kappa_1\sqrt{nT}}\|X(\tau - \beta^*)\|\right) \geq \bar{c}. \tag{6.2}$$

Proof. Fix p and write for brevity $\psi_n = \psi_{n,p}$ where it causes no ambiguity. Throughout this proof we set $x^{1/\infty} = 1$ for any $x \geq 0$. We consider first the case $T \leq \log(eM/s)$. Set $\mathbf{0} = (0, \dots, 0) \in \mathbb{R}^T$, $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^T$. Define the set of vectors

$$\Omega = \left\{\omega \in \mathbb{R}^{MT} \,:\, \omega^j \in \{\mathbf{0}, \mathbf{1}\},\ j = 1, \dots, M, \text{ and } M(\omega) \leq s\right\},$$

and its dilation

$$\mathcal{C}(\Omega) = \left\{\gamma\psi_{n,p}\omega/s^{1/p} \,:\, \omega \in \Omega\right\},$$

where $\gamma > 0$ is an absolute constant to be chosen later. Note that $\mathcal{C}(\Omega) \subset GS(s, M, T)$.

For any ω, ω' in Ω we have $M(\omega - \omega') \leq 2s$. Thus, for $\beta = \gamma\psi_{n,p}\omega/s^{1/p},\ \beta' = \gamma\psi_{n,p}\omega'/s^{1/p}$ parts (a) and (b) of Assumption 6.1 imply respectively

$$\frac{1}{n}\|X\beta - X\beta'\|^2 \geq \frac{\kappa_1^2\gamma^2\psi_{n,p}^2\rho(\omega,\omega')T}{s^{2/p}}, \tag{6.3}$$

$$\frac{1}{n}\|X\beta - X\beta'\|^2 \leq \frac{\kappa_2^2\gamma^2\psi_{n,p}^2\rho(\omega,\omega')T}{s^{2/p}} \tag{6.4}$$

where $\rho(\omega,\omega') = \sum_{j=1}^M I\{\omega^j \neq (\omega')^j\}$ and $I\{\cdot\}$ denotes the indicator function. This and the definition of $\psi_{n,p}$ yield that if part (a) of Assumption 6.1 holds, then for all $\omega, \omega' \in \Omega$ we have

$$\frac{1}{nT}\|X\beta - X\beta'\|^2 \geq \gamma^2\frac{\kappa_1^2\sigma^2}{\kappa_2^2 n}\left(1 + \frac{\log(eM/s)}{T}\right)\rho(\omega,\omega'). \tag{6.5}$$

Also, by definition of β, β',

$$\frac{1}{\sqrt{T}}\|\beta - \beta'\|_{2,p} = \frac{\gamma\sigma}{\kappa_2\sqrt{n}}\left(1 + \frac{\log(eM/s)}{T}\right)^{1/2}(\rho(\omega,\omega'))^{1/p} I\{\omega \neq (\omega')\}. \tag{6.6}$$

For $\theta \in \mathbb{R}^N$, we denote by P_θ the probability distribution of $\mathcal{N}(\theta, \sigma^2 I_{N\times N})$ Gaussian random vector. We denote by $\mathcal{K}(P,Q)$ the Kullback-Leibler divergence between the probability measures P and Q. Then, under part (b) of Assumption 6.1,

$$
\begin{aligned}
\mathcal{K}(P_{X\beta}, P_{X\beta'}) &= \frac{1}{2\sigma^2}\|X\beta - X\beta'\|^2 \\
&\leq \frac{\kappa_2^2\gamma^2}{2\sigma^2 s^{2/p}}\, n\psi_{n,p}^2 \rho(\omega,\omega')T \\
&\leq \gamma^2 s[T + \log(eM/s)] \\
&\leq 2\gamma^2 s\log(eM/s) \qquad (6.7)
\end{aligned}
$$

where we used that $\rho(\omega,\omega') \leq 2s$ for all $\omega,\omega' \in \Omega$. Lemma 8.3 in [32] guarantees the existence of a subset $\mathcal{N}$ of Ω such that

$$
\begin{aligned}
\log(|\mathcal{N}|) &\geq \tilde{c}s\log\left(\frac{eM}{s}\right) \qquad (6.8) \\
\rho(\omega,\omega') &\geq s/4, \forall \omega,\omega' \in \mathcal{N},\ \omega \neq \omega',
\end{aligned}
$$

for some absolute constant $\tilde{c} > 0$, where $|\mathcal{N}|$ denotes the cardinality of $\mathcal{N}$. Combining this with (6.5) and (6.6) we find that the finite set of vectors $\mathcal{C}(\mathcal{N})$ is such that, for all $\beta,\beta' \in \mathcal{C}(\mathcal{N})$, $\beta \neq \beta'$,

$$
\frac{1}{\sqrt{T}}\|\beta - \beta'\|_{2,p} \geq \frac{\gamma\sigma s^{1/p}}{4^{1/p}\kappa_2\sqrt{n}}\left(1 + \frac{\log(eM/s)}{T}\right)^{1/2} = \frac{\gamma}{4^{1/p}}\psi_{n,p}\,,
$$

and under part (a) of Assumption 6.1,

$$
\frac{1}{nT}\|X\beta - X\beta'\|^2 \geq \gamma^2\frac{\kappa_1^2\sigma^2 s}{4\kappa_2^2 n}\left(1 + \frac{\log(eM/s)}{T}\right) = \frac{\gamma^2}{4}\kappa_1^2\psi_{n,2}^2\,.
$$

Furthermore, by (6.7) and (6.8) for all $\beta,\beta' \in \mathcal{C}(\mathcal{N})$ under part (b) of Assumption 6.1 we have

$$
\mathcal{K}(P_{X\beta}, P_{X\beta'}) \leq \frac{1}{16}\log\left(|\mathcal{N}|\right) = \frac{1}{16}\log\left(|\mathcal{C}(\mathcal{N})|\right)
$$

for an absolute constant $\gamma > 0$ chosen small enough. Thus, the result follows by application of Theorem 2.7 in [35].

Consider now the case $T > \log(eM/s)$. Introduce the set of vectors

$$
\Omega' = \left\{\omega \in \mathbb{R}^{MT} : \omega = (\omega^1,\ldots,\omega^M),\ \omega^j \in \{0,1\}^T \text{ if } j \leq s \text{ and } \omega^j = \mathbf{0} \text{ otherwise}\right\},
$$

and the associated dilated set $\mathcal{C}(\Omega')$ defined as above. Note that $\mathcal{C}(\Omega') \subset GS(s,M,T)$.

For any $\omega,\omega' \in \Omega'$ we define $\rho'(\omega,\omega') = \sum_{j=1}^M\sum_{t=1}^T I\{\omega_{tj} \neq \omega'_{tj}\} = \sum_{j=1}^s\sum_{t=1}^T I\{\omega_{tj} \neq \omega'_{tj}\}$.

We assume first that $T \geq 16$ and $s \geq 16$. Then Varshamov-Gilbert Lemma (see Lemma 2.9 in [35]) guarantees that there exists a subset $\mathcal{N}'$ of Ω' such that

$$
\begin{aligned}
|\mathcal{N}'| &\geq 2^{Ts/8}, \qquad (6.9) \\
\rho'(\omega,\omega') &\geq \frac{Ts}{8}, \forall \omega,\omega' \in \mathcal{N}',\ \omega \neq \omega'.
\end{aligned}
$$

Next for any $\omega, \omega' \in \mathcal{N}'$ we have $M(\omega - \omega') \leq 2s$, and thus under parts (a) and (b) of Assumption 6.1 we have, respectively,

$$\frac{1}{n}\|X\beta - X\beta'\|^2 \geq \frac{\kappa_1^2\gamma^2\psi_n^2\rho'(\omega,\omega')}{s^{2/p}}, \qquad \frac{1}{n}\|X\beta - X\beta'\|^2 \leq \frac{\kappa_2^2\gamma^2\psi_n^2\rho'(\omega,\omega')}{s^{2/p}}$$

where $\beta = \gamma\psi_n\omega/s^{1/p},\ \beta' = \gamma\psi_n\omega'/s^{1/p}$ are any two elements of $\mathcal{C}(\mathcal{N}')$.

Now, using Lemma A.3 in the Appendix we get that, for all $\omega, \omega' \in \mathcal{N}'$ such that $\omega \neq \omega'$,

$$\|\omega - \omega'\|_{2,p} \geq \left(\frac{s}{16}\right)^{1/p}\frac{\sqrt{T}}{4}, \quad \forall\ 1 \leq p \leq \infty. \tag{6.10}$$

Thus, for all $\beta, \beta' \in \mathcal{C}(\mathcal{N}')$ such that $\beta \neq \beta'$ we have

$$\frac{1}{\sqrt{T}}\|\beta - \beta'\|_{2,p} \ = \ \frac{\gamma\psi_n}{s^{1/p}\sqrt{T}}\|\omega - \omega'\|_{2,p} \geq \frac{\gamma}{16^{1/p}4}\psi_n$$

(recall that $\psi_n = \psi_{n,p}$), and under part (a) of Assumption 6.1,

$$\frac{1}{nT}\|X\beta - X\beta'\|^2 \ \geq \ \frac{\gamma^2}{8}\frac{s\kappa_1^2\sigma^2}{\kappa_2^2 n}\left(1 + \frac{\log(eM/s)}{T}\right) = \frac{\gamma^2}{8}\kappa_1^2\psi_{n,2}^2.$$

Furthermore, for all $\beta, \beta' \in \mathcal{C}(\mathcal{N}')$ under part (b) of Assumption 6.1,

$$\mathcal{K}(P_{X\beta}, P_{X\beta'}) \ \leq \ 2\gamma^2 sT \leq \frac{1}{16}\log(|C(\mathcal{N}')|),$$

where, in view of (6.9), the last inequality holds for an absolute constant $\gamma > 0$ chosen small enough. We apply again Theorem 2.7 in [35] to get the result.

Finally, if $T > \log(eM/s)$ and $T < 16$, $s < 16$, then the rate ψ_n is of the order $1/n$. This is the standard parametric rate and the lower bounds are easily obtained by reduction to distinguishing between two elements of $GS(s, M, T)$. ∎

As a consequence of Theorem 6.1, we get, for example, the lower bounds for the squared loss $\ell(u) = u^2$ and for the indicator loss $\ell(u) = I\{u \geq 1\}$. The indicator loss is relevant for comparison with the upper bounds of Corollaries 4.1 and 5.2. For example, Theorem 6.1 with this loss and $p = 1, 2$ implies that there exists $\beta^* \in GS(s, M, T)$ such that, for any estimator τ of β^*,

$$\frac{1}{\sqrt{nT}}\|X(\tau - \beta^*)\| \geq C\sqrt{\frac{s}{n}}\left(1 + \frac{\log(eM/s)}{T}\right)^{1/2}$$

and

$$\frac{1}{\sqrt{T}}\|\tau - \beta^*\| \geq C\sqrt{\frac{s}{n}}\left(1 + \frac{\log(eM/s)}{T}\right)^{1/2}, \qquad \frac{1}{\sqrt{T}}\|\tau - \beta^*\|_{2,1} \geq C\frac{s}{\sqrt{n}}\left(1 + \frac{\log(eM/s)}{T}\right)^{1/2}$$

with a positive probability (independent of n, s, M, T) where $C > 0$ is some constant. The rate on the right-hand side of these inequalities is of the same order as in the corresponding upper bounds in Corollary 4.1, modulo that $\log M$ is replaced here by $\log(eM/s)$. We conjecture that the factor $\log(eM/s)$ and not $\log M$ corresponds to the optimal rate; actually, we know that this conjecture is true when $T = 1$ and the risk is defined by the prediction error with $\ell(u) = u^2$ [32].

A weaker version of Theorem 6.1, with $\ell(u) = u^2$, $p = 2$ and suboptimal rate of the order $[s\log(M/s)/(nT)]^{1/2}$ is established in [17].

Remark 6.1. *For the model with usual (non-grouped) sparsity, which corresponds to $T = 1$, the set $GS(s, M, 1)$ coincides with the ℓ_0-ball of radius s in $\mathbb{R}^M$. Therefore, Theorem 6.1 generalizes the minimax lower bounds on ℓ_0-balls recently obtained in [30] and [32] for the usual sparsity model. Those papers considered only the prediction error and the ℓ_2 error under the squared loss $\ell(u) = u^2$. Theorem 6.1 covers any ℓ_p error with $1 \le p \le \infty$ and applies with general loss functions $\ell(\cdot)$. As a particular instance, for the indicator loss $\ell(u) = I\{u \ge 1\}$ and $T = 1$, the lower bounds of Theorem 6.1 show that the upper bounds for the prediction error and the ℓ_p errors ($1 \le p \le \infty$) of the usual Lasso estimator established in [4] and [21] cannot be improved in a minimax sense on ℓ_0-balls up to logarithmic factors. Note that this conclusion cannot be deduced from the lower bounds of [30] and [32].*

7 Lower bounds for the Lasso

In this section we establish lower bounds on the prediction and estimation accuracy of the Lasso estimator. As a consequence, we can emphasize the advantages of using the Group Lasso estimator as compared to the usual Lasso in some important particular cases.

The Lasso estimator is a solution of the minimization problem

$$\min_{\beta \in \mathbb{R}^K} \frac{1}{N} \|X\beta - y\|^2 + 2r\|\beta\|_1, \tag{7.1}$$

where $\|\beta\|_1 = \sum_{j=1}^K |\beta_j|$ and r is a positive parameter. The following notations apply only to this section. For any vector $\beta \in \mathbb{R}^K$ and any subset $J \subseteq \mathbb{N}_K$, we denote by $\beta_{|J}$ the vector in $\mathbb{R}^K$ which has the same coordinates as β on J and zero coordinates on the complement J^c of J, $J'(\beta) = \{j : \beta_j \neq 0\}$ and $M'(\beta) = |J'(\beta)|$.

We will use the following standard assumption on the matrix X (the Restricted Eigenvalue condition in [4]).

Assumption 7.1. *Fix $s' \ge 1$. There exists a positive number κ' such that*

$$\min\left\{ \frac{\|X\Delta\|}{\sqrt{N}\|\Delta_{|J}\|} \;:\; |J| \le s', \Delta \in \mathbb{R}^K \setminus \{0\}, \sum_{j \in J^c} |\Delta_j| \le 3 \sum_{j \in J} |\Delta_j| \right\} \ge \kappa',$$

where J^c denotes the complement of the set of indices J.

Theorem 7.1. *Let Assumption 7.1 be satisfied. Assume that $W \in \mathbb{R}^N$ is a random vector with i.i.d. $\mathcal{N}(0, \sigma^2)$ gaussian components, $\sigma^2 > 0$. Set $r = A\sigma\sqrt{\frac{\phi \log K}{N}}$ where $A > 2\sqrt{2}$ and ϕ is the maximal diagonal element of the matrix $\Psi = \frac{1}{N}X^\top X$. If $\hat{\beta}^L$ is a solution of problem* (7.1)*, then with probability at least $1 - K^{1-\frac{A^2}{8}}$ we have*

$$\frac{1}{N}\|X(\hat{\beta}^L - \beta^*)\|^2 \;\ge\; M'(\hat{\beta}^L)\frac{A^2\sigma^2\phi \log K}{4\phi_{\max} N}, \tag{7.2}$$

$$\|\hat{\beta}^L - \beta^*\| \;\ge\; \frac{A\sigma}{2\phi_{\max}}\sqrt{M'(\hat{\beta}^L)\frac{\phi \log K}{N}}, \tag{7.3}$$

where $\phi_{\max}$ is the maximum eigenvalue of the matrix Ψ. If, in addition, $M'(\beta^*) \le s'$, and

$$\min\{|\Psi_{jj}\beta_j^*| : j \in \mathbb{N}_m,\ \beta_j^* \ne 0\} > \left(\frac{3}{2} + \frac{16s'}{\kappa'^2}\max_{j\ne k}|\Psi_{jk}|\right) r, \tag{7.4}$$

where Ψ_{jk} denotes the (j,k)-th entry of matrix Ψ, then with the same probability we have

$$M'(\hat\beta^L) \ge M'(\beta^*). \tag{7.5}$$

Proof. Inequality (B.3) in [4] yields (7.2) on the event $\mathcal{A} = \left\{\frac{1}{N}\|X^\top W\|_\infty \le \frac{r}{2}\right\}$ of probability $\mathbb{P}(\mathcal{A}) \ge 1 - K^{1-\frac{A^2}{8}}$.

Next, (7.3) follows from (7.2) and the inequality

$$\frac{1}{N}(\hat\beta^L - \beta^*)^\top X^\top X(\hat\beta^L - \beta^*) \le \phi_{\max}\|\hat\beta^L - \beta^*\|^2.$$

We now prove (7.5). If $M'(\hat\beta^L) < M'(\beta^*)$ then there exists $j \in J'(\hat\beta^L)^c \cap J'(\beta^*)$. Set $\Delta = \beta^* - \hat\beta^L$ and recall that $\Psi = \frac{1}{N}X^\top X$. Using that any Lasso solution $\hat\beta^L$ satisfies

$$\begin{cases} \frac{1}{N}(X^\top(y - X\hat\beta^L))_j = \operatorname{sign}(\hat\beta_j^L) r, & \text{if } \hat\beta_j^L \ne 0, \\ \left|\frac{1}{N}(X^\top(y - X\hat\beta^L))_j\right| \le r, & \text{if } \hat\beta_j^L = 0. \end{cases} \tag{7.6}$$

and the triangle inequality we get, on the event $\mathcal{A}$, that $|(\Psi\Delta)_j| \le \frac{3r}{2}$. Consequently,

$$|\Psi_{jj}\beta_j^*| = |\Psi_{jj}\Delta_j| = \left|(\Psi\Delta)_j - \sum_{k\ne j}\Psi_{jk}\Delta_k\right| \le \frac{3r}{2} + \|\Delta\|_1 \max_{j\ne k}|\Psi_{jk}|. \tag{7.7}$$

Next, Corollary B.2 in [4] yields that, on the event $\mathcal{A}$,

$$\|\Delta_{|J'(\beta^*)^c}\|_1 \quad\le\quad 3\|\Delta_{|J'(\beta^*)}\|_1.$$

Thus, the Cauchy-Schwarz inequality, Assumption 7.1 and [4, Inequality (7.8)] give that, on the event $\mathcal{A}$,

$$\|\Delta\|_1 \quad\le\quad 4\|\Delta_{|J'(\beta^*)}\|_1 \le 4\sqrt{s'}\|\Delta_{|J'(\beta^*)}\| \le \frac{4\sqrt{s'}}{\kappa'}(\Delta^\top\Psi\Delta)^{1/2} \le \frac{16s'}{\kappa'^2} r. \tag{7.8}$$

Combining (7.7) and (7.8) yields, on the event $\mathcal{A}$, that

$$|\Psi_{jj}\beta_j^*| \le \left(\frac{3}{2} + \frac{16s'}{\kappa'^2}\max_{j\ne k}|\Psi_{jk}|\right) r,$$

which contradicts the condition (7.4).

■

Let us emphasize that the Theorem 7.1 establishes lower bounds, which hold for every Lasso solution if $\hat{\beta}_L$ is not unique.

Theorem 7.1 highlights several limitations of the usual Lasso as compared to the Group Lasso. Let us explain this point in the multi-task learning case. There, the usual Lasso estimator $\hat{\beta}^L$ is a solution of the following optimization problem

$$\min\left\{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\|X_t\beta_t - y_t\|^2 + 2r\sum_{t=1}^{T}\sum_{j=1}^{M}|\beta_{tj}|\right\}.$$

By comparing the prediction error lower bound in Theorem 7.1 for this estimator with the corresponding upper bound for Group Lasso estimator derived in Corollary 4.1, we reach the following conclusions.

- *The usual Lasso does not enjoy any dimension independence phenomenon as compared to the Group Lasso.*

  In the multi-task learning setting we have $N = nT$, $K = MT$. Assume that the tasks' design matrices are orthogonal, namely $X_t^\top X_t/n = I_{M\times M}$ for every $t \in \mathbb{N}_T$. Hence, $\Psi = I_{TM\times TM}/T$, so that $\phi_{\max} = \phi = 1/T$ and $\Psi_{jj} = 1/T$ for all j. Let a special instance of group sparsity assumption be realized, namely, all vectors β_t^* have exactly s non-zero entries at the same positions. Then, $M(\beta^*) = s$ and $M'(\beta^*) = sT$. Moreover, condition (7.4) simplifies to the requirement that

  $$\min_{j:\beta_j^*\neq 0}|\beta_j^*| \geq \frac{3A\sigma}{2}\sqrt{\frac{\log(MT)}{n}}.$$

  We conclude by inequalities (7.2) and (7.5) that, with probability at least $1-(MT)^{1-\frac{A^2}{8}}$,

  $$\frac{1}{nT}\|X(\hat{\beta}^L - \beta^*)\|^2 \geq A^2\sigma^2 s\frac{\log(MT)}{4n}. \tag{7.9}$$

  This bound holds no matter what the number of tasks T is. In contrast, the bounds in Corollary 4.1 can be made independent of the dimension M and of the number of tasks T as soon as $T \geq \log M$. Specifically, under the above assumptions we have, recalling Definition 4.1, that $\kappa_{\mathrm{MT}} \geq 1$ and by (4.3), with probability close to 1, every Group Lasso solution $\hat{\beta}$ satisfies

  $$\frac{1}{nT}\|X(\hat{\beta} - \beta^*)\|^2 \leq 128\sigma^2\frac{s}{n}\left(1 + \frac{A\log M}{T}\right). \tag{7.10}$$

- *The Group Lasso achieves faster rates of convergence in some cases as compared to the usual Lasso.* We consider separately two cases. The first one is already discussed the preceding remark. It corresponds to $T \geq \log M$. Then the upper bound for the Group Lasso (7.10) is smaller than the lower bound (7.9) for the Lasso by a logarithmic factor. This factor can be large if T is large, for example exponential in n, so that (7.9) gives no convergence result for the Lasso. The second case is $T < \log M$. Then the lower bound (7.9) is of the order $s(\log M)/n$, while the upper bound (7.10) is of the order $s(\log M)/(nT)$. The ratio is of the order T in favor of the Group Lasso.

In (7.9) and (7.10) we have only compared the prediction errors of the two estimators. In view of inequality (4.6) and Theorem 7.1, similar observations are valid for the ℓ_2 estimation errors.

8 Non-Gaussian noise

In this section, we show that the above results extend to non-gaussian noise. We consider here the multi-task setting described in Section 2.1 and we only assume that the components of random vector W are independent with zero mean and finite fourth moment $\mathbb{E}[W_{tj}^4]$. As we shall see the results remain similar to those of the previous sections, though the concentration effect is weaker.

We need the following technical assumption.

Assumption 8.1. *The matrix X is such that*

$$\max_{t\in\mathbb{N}_T}\left(\frac{1}{n}\sum_{i=1}^{n}\max_{j\in\mathbb{N}_M}|(x_{ti})_j|^2\right)\leq x_*^2$$

for a finite constant x_*.

This assumption is quite mild. It is satisfied for example, if all $(x_{ti})_j$ are bounded in absolute value by a constant uniformly in i,t,j. We have the two following theorems.

Theorem 8.1. *Consider the model* (2.1) *for any $M\geq 2$, $T,n\geq 1$. Assume that the components of random vector W are independent with zero mean, $\max_{t\in\mathbb{N}_T,j\in\mathbb{N}_M}\mathbb{E}[W_{tj}^4]\leq b^4$, all diagonal elements of the matrix $X^\top X/n$ are equal to 1 and $M(\beta^*)\leq s$. Let also Assumption 8.1 be satisfied. Set*

$$\lambda=\frac{x_*b}{\sqrt{nT}}\left(1+\frac{(\log M)^{3/2+\delta}}{\sqrt{T}}\right)^{1/2},$$

with $\delta>0$. Then with probability at least $1-\frac{4\sqrt{\log(2M)}[(8\log(12M))^2+1]^{1/2}}{(\log M)^{3/2+\delta}}$, for any solution $\hat\beta$ of problem (2.6) *we have*

$$\frac{1}{nT}\|X(\hat\beta-\beta^*)\|^2\ \leq\ \frac{4x_*b}{\sqrt{nT}}\left(1+\frac{(\log M)^{3/2+\delta}}{\sqrt{T}}\right)^{1/2}\|\beta^*\|_{2,1}. \tag{8.1}$$

If, in addition, Assumption 4.1 holds, then with the same probability for any solution $\hat\beta$ of problem (2.6) *we have*

$$\frac{1}{nT}\|X(\hat\beta-\beta^*)\|^2\leq\frac{16x_*^2b^2}{\kappa_{\mathrm{MT}}^2}\frac{s}{n}\left(1+\frac{(\log M)^{3/2+\delta}}{\sqrt{T}}\right), \tag{8.2}$$

$$\frac{1}{\sqrt{T}}\|\hat\beta-\beta^*\|_{2,1}\leq\frac{16x_*b}{\kappa_{\mathrm{MT}}^2}\frac{s}{\sqrt{n}}\left(1+\frac{(\log M)^{3/2+\delta}}{\sqrt{T}}\right)^{1/2}, \tag{8.3}$$

$$M(\hat\beta)\leq\frac{64\phi_{\mathrm{MT}}}{\kappa_{\mathrm{MT}}^2}s, \tag{8.4}$$

where ϕ_{MT} is the largest eigenvalue of the matrix $X^\top X/n$. If, in addition, $\kappa_{\mathrm{MT}}(2s)>0$, then with the same probability for any solution $\hat\beta$ of problem (2.6) *we have*

$$\frac{1}{\sqrt{T}}\|\hat\beta-\beta^*\|\leq\frac{4\sqrt{10}\,x_*b}{\kappa^2(2s)}\sqrt{\frac{s}{n}}\left(1+\frac{(\log M)^{3/2+\delta}}{\sqrt{T}}\right)^{1/2}.$$

Theorem 8.2. *Consider the model (2.1) for $M \geq 2$, $T, n \geq 1$. Let the assumptions of Theorem 8.1 be satisfied and let Assumption 5.1 hold with the same s. Set*

$$\tilde{c} = \left(\frac{3}{2} + \frac{8}{7(\alpha - 1)}\right) x_* b.$$

Let λ be as in Theorem 8.1. Then with probability at least $1 - \frac{4\sqrt{\log(2M)}[(8\log(12M))^2+1]^{1/2}}{(\log M)^{3/2+\delta}}$, for any solution $\hat{\beta}$ of problem (2.6) we have

$$\frac{1}{\sqrt{T}}\|\hat{\beta} - \beta^*\|_{2,\infty} \leq \frac{\tilde{c}}{\sqrt{n}}\left(1 + \frac{(\log M)^{3/2+\delta}}{\sqrt{T}}\right)^{1/2}.$$

If, in addition, it holds that

$$\min_{j \in J(\beta^*)} \frac{1}{\sqrt{T}}\|(\beta^*)^j\| > \frac{2\tilde{c}}{\sqrt{n}}\left(1 + \frac{(\log M)^{3/2+\delta}}{\sqrt{T}}\right)^{1/2},$$

then with the same probability for any solution $\hat{\beta}$ of problem (2.6) the set of indices

$$\hat{J} = \left\{j : \frac{1}{\sqrt{T}}\|\hat{\beta}^j\| > \frac{\tilde{c}}{\sqrt{n}}\left(1 + \frac{(\log M)^{3/2+\delta}}{\sqrt{T}}\right)^{1/2}\right\}$$

estimates correctly the sparsity pattern $J(\beta^*)$:

$$\hat{J} = J(\beta^*).$$

Proof. The proofs of these theorems are similar to those of Theorems 3.1 and 5.1 up to a modification of the bound on $\mathbb{P}(\mathcal{A}^c)$ in Lemma 3.1. We consider now the event

$$\mathcal{A} = \left\{\max_{j=1}^{M} \sqrt{\sum_{t=1}^{T}\left(\sum_{i=1}^{n}(x_{ti})_j W_{ti}\right)^2} \leq \lambda n T\right\}.$$

Define the random variables

$$Y_{tj} = \left(\sum_{i=1}^{n}(x_{ti})_j W_{ti}\right)^2 - \sum_{i=1}^{n}|(x_{ti})_j|^2 \mathbb{E}[W_{ti}^2], \;\; j = 1, \dots, M, \; t = 1, \dots, T.$$

We have

$$\begin{aligned}
\mathbb{P}(\mathcal{A}^c) &= \mathbb{P}\left(\max_{1\leq j\leq M}\sum_{t=1}^{T}\left(\sum_{i=1}^{n}(x_{ti})_j W_{ti}\right)^2 \geq (\lambda n T)^2\right) \\
&\leq \mathbb{P}\left(\max_{1\leq j\leq M}\sum_{t=1}^{T} Y_{tj} \geq x_*^2 b^2 n\sqrt{T}(\log M)^{3/2+\delta}\right) \\
&\leq \frac{\mathbb{E}\,\max_{1\leq j\leq M}\left|\sum_{t=1}^{T} Y_{tj}\right|}{x_*^2 b^2 n\sqrt{T}(\log M)^{3/2+\delta}}.
\end{aligned}$$

Applying the maximal moment inequality of Lemma 9.1 below with $m = 1$ and constant $c(1) = 2$ we obtain

$$
\begin{aligned}
\mathbb{E} \max_{1 \le j \le M} \left| \sum_{t=1}^{T} Y_{tj} \right| &\le \sqrt{8 \log(2M)}\ \mathbb{E} \left(\left[\sum_{t=1}^{T} \max_{1 \le j \le M} Y_{tj}^2 \right]^{1/2} \right) \qquad (8.5) \\
&\le \sqrt{8 \log(2M)}\ \left[\sum_{t=1}^{T} \mathbb{E} \left(\max_{1 \le j \le M} Y_{tj}^2 \right) \right]^{1/2} \\
&\le 4\sqrt{\log(2M)}\ \left\{ b^4 x_*^4 n^2 T + \sum_{t=1}^{T} \mathbb{E} \left(\max_{1 \le j \le M} \left| \sum_{i=1}^{n} (x_{ti})_j W_{ti} \right|^4 \right) \right\}^{1/2} .
\end{aligned}
$$

By the maximal moment inequality of Lemma 9.1 with $m = 4$ and constant $c(4) = 12$ (since $M \ge 2$) the last expectation is bounded, for any $t = 1, \dots, T$, as

$$
\mathbb{E} \left(\max_{1 \le j \le M} \left| \sum_{i=1}^{n} (x_{ti})_j W_{ti} \right|^4 \right) \le (8 \log(12M))^2 \mathbb{E} \left(\left[\sum_{i=1}^{n} \max_{1 \le j \le M} (x_{ti})_j^2 W_{ti}^2 \right]^2 \right) .
$$

Setting for brevity $\overline{x}_i = \max_{1 \le j \le M} (x_{ti})_j^2$ we have

$$
\begin{aligned}
\mathbb{E} \left(\left[\sum_{i=1}^{n} \max_{1 \le j \le M} (x_{ti})_j^2 W_{ti}^2 \right]^2 \right) &\le b^4 \left(\sum_{i \ne k} \overline{x}_i \overline{x}_k + \sum_{i=1}^{n} \overline{x}_i^2 \right) \\
&= b^4 \left(\sum_{i=1}^{n} \overline{x}_i \right)^2 \le b^4 x_*^4 n^2 .
\end{aligned}
$$

Combining the above four displays yields

$$
\mathbb{P}(\mathcal{A}^c) \le \frac{4\sqrt{\log(2M)} \left[(8 \log(12M))^2 + 1 \right]^{1/2}}{(\log M)^{3/2+\delta}} .
$$

■

9 Maximal moment inequality

In this section we prove the following inequality for the m-th moment of maxima of sums of independent random variables.

Lemma 9.1. (Maximal moment inequality) *Let $Z_1, \dots, Z_n$ be independent random vectors in $\mathbb{R}^M$, and let $Z_{i,j}$ denote the j-th component of Z_i. Then for any $m \ge 1$ and $M \ge 1$ we have*

$$
\mathbb{E} \left(\max_{1 \le j \le M} \left| \sum_{i=1}^{n} \left(Z_{i,j} - \mathbb{E} Z_{i,j} \right) \right|^m \right) \le \left[8 \log(c(m) M) \right]^{m/2} \mathbb{E} \left(\left[\max_{1 \le j \le M} \sum_{i=1}^{n} Z_{i,j}^2 \right]^{m/2} \right) ,
$$

where $c(m) = \min\{c > 0 : \, e^{m-1} - 1 \le (c-2)M\}$. In particular, $2 \le c(m) \le e^{m-1} + 1$.

Before giving the proof, we make some comments. The case $m = 2$ of Lemma 9.1 implies – modulo constants – Nemirovski's inequality (see [27], page 188, and [13], Corollary 2.4). In general, Nemirovski's inequality concerns the second moment of ℓ_p-norms ($1 \leq p \leq \infty$) of sums of independent random variables in $\mathbb{R}^M$, whereas we only consider $p = \infty$. On the other hand, even for $m = 2$ Lemma 9.1 is more general than what is given by Nemirovski's inequality because we interchange the maximum and the sum on the right hand side. The case $M = 1$ of Lemma 9.1 yields the Marcinkiewicz-Zygmund inequality (see [29], page 82), and as an immediate consequence the inequality

$$\mathbb{E}\left(\Big|\sum_{i=1}^n \xi_i\Big|^m\right) \leq [8\log(c(m))]^{m/2} n^{m/2-1} \sum_{i=1}^n \mathbb{E}\,|\xi_i|^m, \quad m \geq 2, \tag{9.1}$$

for independent zero-mean random variables ξ_i. Thus, as a particular instance, we give a short proof of (9.1) and provide the explicit constant. This constant is of the optimal order in m but larger than the one obtainable from the recent sharp moment inequality due to Rio [33].

Proof. Let $(\varepsilon_1, \ldots, \varepsilon_n)$ be a sequence of i.i.d. Rademacher random variables independent of $\mathbf{Z} = (Z_1, \ldots, Z_n)$. Let $\mathbb{E}_{\mathbf{Z}}$ denote conditional expectation given $\mathbf{Z}$. By Hoeffding's inequality, for all $L > 0$ and all i and j,

$$\mathbb{E}_{\mathbf{Z}} \exp[Z_{i,j}\varepsilon_i/L] \leq \exp[Z_{i,j}^2/(2L^2)]. \tag{9.2}$$

Define

$$\zeta = \max_{1\leq j\leq M}\Big|\sum_{i=1}^n Z_{i,j}\varepsilon_i\Big|.$$

Using successively Jensen's inequality (the function $x \mapsto \log^m(x + e^{m-1} - 1)$ is concave for $x \geq 1$), the inequality $e^{|x|} \leq e^x + e^{-x}$, $\forall\, x \in \mathbb{R}$, the independence of ε_i, and (9.2), we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}(\zeta^m) &\leq L^m \mathbb{E}_{\mathbf{Z}} \log^m\Big\{\exp[\zeta/L] + e^{m-1} - 1\Big\} \\
&\leq L^m \log^m\Big\{\mathbb{E}_{\mathbf{Z}} \exp[\zeta/L] + e^{m-1} - 1\Big\} \\
&\leq L^m \log^m\Big\{\sum_{j=1}^M \mathbb{E}_{\mathbf{Z}} \exp\Big[\Big|\sum_{i=1}^n Z_{i,j}\varepsilon_i\Big|/L\Big] + e^{m-1} - 1\Big\} \\
&\leq L^m \log^m\left\{2M \exp\Big[\max_{1\leq j\leq M}\sum_{i=1}^n Z_{i,j}^2/(2L^2)\Big] + e^{m-1} - 1\right\}.
\end{aligned}$$

Note that $2Mx + e^{m-1} - 1 \leq c(m)Mx$ for all $x \geq 1$, where $c(m)$ is the constant defined in the statement of the lemma. This and the previous display yield

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}(\zeta^m) &\leq L^m \log^m\left\{c(m)M \exp\Big[\max_{1\leq j\leq M}\sum_{i=1}^n Z_{i,j}^2/(2L^2)\Big]\right\} \\
&= L^m \left\{\log(c(m)M) + \frac{\max_{1\leq j\leq M}\sum_{i=1}^n Z_{i,j}^2}{2L^2}\right\}^m.
\end{aligned}$$

Choosing

$$L = \sqrt{\frac{\max_{1\le j\le M}\sum_{i=1}^n Z_{i,j}^2}{2\log(c(m)M)}}$$

gives

$$\mathbb{E}_{\mathbf{Z}}\left(\max_{1\le j\le M}\left|\sum_{i=1}^n Z_{i,j}\varepsilon_i\right|^m\right) \le \left[2\log(c(m)M)\max_{1\leqslant j\le M}\sum_{i=1}^n Z_{i,j}^2\right]^{m/2}.$$

Hence,

$$\mathbb{E}\left(\max_{1\le j\le M}\left|\sum_{i=1}^n Z_{i,j}\varepsilon_i\right|^m\right) \le \left[2\log(c(m)M)\right]^{m/2}\mathbb{E}\left(\left[\max_{1\le j\le M}\sum_{i=1}^n Z_{i,j}^2\right]^{m/2}\right).$$

Finally, we de-symmetrize (see Lemma 2.3.1 page 108 in [37]):

$$\left(\mathbb{E}\max_{1\le j\le M}\left|\sum_{i=1}^n (Z_{i,j}-\mathbb{E}Z_{i,j})\right|^m\right)^{1/m} \le 2\left(\mathbb{E}\max_{1\le j\le M}\left|\sum_{i=1}^n Z_{i,j}\varepsilon_i\right|^m\right)^{1/m}.$$

■

Acknowledgments

Part of this work was supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778 as well as by the EPSRC Grant EP/D071542/1 and ANR "Parcimonie".

A Auxiliary results

Here we collect some auxiliary results which we have use in the paper.

The first result is taken from [9, Eq. (27)] and was used in the proof of Lemma 3.1.

Lemma A.1. *Let* $\xi_1,\dots,\xi_N$ *be i.i.d.* $\mathcal{N}(0,1)$, $v=(v_1,\dots,v_N)\neq 0$, $\eta_v = \frac{1}{\sqrt{2}\|v\|}\sum_{i=1}^N(\xi_i^2-1)v_i$ *and* $m(v)=\frac{\|v\|_\infty}{\|v\|}$. *We have, for all* $x>0$, *that*

$$\mathbb{P}(|\eta_v|>x)\le 2\exp\left(-\frac{x^2}{2(1+\sqrt{2}xm(v))}\right).$$

The next lemma provides the link between Assumptions 5.1 and 3.1 and was used extensively in our analysis in Section 5.

Lemma A.2. *Let Assumption 5.1 be satisfied. Then Assumption 3.1 is satisfied with* $\kappa=\sqrt{(1-1/\alpha)\phi}$.

Proof. We use here the notations introduced in the proof of Theorem 5.1. For any subset J of $\mathbb{N}_M$ such that $|J| \le s$ and any $\Delta \in \mathbb{R}^K$ we have

$$
\begin{aligned}
|\Delta_J^\top(\Psi - \phi I_{K\times K})\Delta_J| &\le \sum_{j,j'\in J}\sum_{t=1}^{K_j}\sum_{t'=1}^{K_{j'}}\left|\left(\tilde{\Psi}[j,j']\right)_{t,t'}\right| |\tilde{\Delta}_t^j||\tilde{\Delta}_{t'}^{j'}| \\
&= \sum_{j,j'\in J}\sum_{t=1}^{\min(K_j,K_{j'}}\left|\left(\tilde{\Psi}[j,j']\right)_{t,t}\right| |\tilde{\Delta}_t^j||\tilde{\Delta}_t^{j'}| \\
&\quad + \sum_{j,j'\in J}\sum_{t=1}^{K_j}\sum_{t'=1,t'\neq t}^{K_{j'}}\left|\left(\tilde{\Psi}[j,j']\right)_{t,t'}\right| |\tilde{\Delta}_t^j||\tilde{\Delta}_{t'}^{j'}|.
\end{aligned}
$$

We now treat separately the first and second terms in the right-hand side of the above display. For the first term we have, using consecutively Assumption 5.1, Cauchy-Schwarz and Minkowski's inequality for the Euclidean norm in $\mathbb{R}^{K_j}$, that

$$
\begin{aligned}
\sum_{j,j'\in J}\sum_{t=1}^{K_j}\left|\left(\tilde{\Psi}[j,j']\right)_{t,t}\right| |\tilde{\Delta}_t^j||\tilde{\Delta}_t^{j'}| &\le \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\sum_{t=1}^{K_j}\left(\sum_{j\in J}|\tilde{\Delta}_t^j|\right)^2 \\
&\le \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\|\Delta_J\|_{2,1}^2 \\
&\le \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}}\|\Delta_J\|^2.
\end{aligned}
$$

For the second term we get, using Assumption 5.1 and Cauchy-Schwarz's inequality twice, that

$$
\begin{aligned}
\sum_{j,j'\in J}\sum_{t=1}^{K_j}\sum_{t'=1,t'\neq t}^{K_{j'}}\left|\left(\tilde{\Psi}[j,j']\right)_{t,t'}\right| |\tilde{\Delta}_t^j||\tilde{\Delta}_{t'}^{j'}| &\le \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\left(\sum_{j\in J}\frac{1}{\sqrt{K_j}}\sum_{t=1}^{K_j}|\Delta_t^j|\right)^2 \\
&\le \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}}\|\Delta_J\|^2.
\end{aligned}
$$

Combining the two above displays yields

$$
\begin{aligned}
\frac{\Delta_J^\top\Psi\Delta_J}{\|\Delta_J\|^2} &= \phi + \frac{\Delta_J^\top(\Psi - \phi I_{K\times K})\Delta_J}{\|\Delta_J\|^2} \\
&\ge \phi\left(1 - \frac{2\lambda_{\min}}{14\alpha\lambda_{\max}}\right).
\end{aligned}
$$

We proceed similarly to treat the quantity $|\Delta_{J^c}\Psi\Delta_J|$. We have, using Assumption 5.1, Cauchy-

Schwarz and Minkowski's inequalities, that

$$
\begin{aligned}
|\Delta_{J^c}\Psi\Delta_J| \leq & \sum_{j\in J^c, j'\in J}\sum_{t=1}^{K_j}\left|\left(\tilde{\Psi}[j,j']\right)_{t,t}\right| |\tilde{\Delta}_t^j||\tilde{\Delta}_t^{j'}| \\
& \qquad + \sum_{j\in J^c, j'\in J}\sum_{t=1}^{K_j}\sum_{t'=1,t'\neq t}^{K_{j'}}\left|\left(\tilde{\Psi}[j,j']\right)_{t,t'}\right| |\tilde{\Delta}_t^j||\tilde{\Delta}_{t'}^{j'}| \\
\leq & \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\|\Delta_{J^c}\|_{2,1}\|\Delta_J\|_{2,1} \\
& \qquad + \frac{\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\left(\sum_{j\in J}\sum_{t=1}^{K_j}\frac{1}{\sqrt{K_j}}|\Delta_t^j|\right)\left(\sum_{j\in J^c}\sum_{t=1}^{K_j}\frac{1}{\sqrt{K_j}}|\Delta_t^j|\right) \\
\leq & \frac{2\lambda_{\min}\phi}{14\alpha\lambda_{\max}s}\|\Delta_J\|_{2,1}\|\Delta_{J^c}\|_{2,1}.
\end{aligned}
$$

Next we have, for any vector $\Delta \in \mathbb{R}^K$ satisfying the inequality $\sum_{j\in J^c}\lambda_j\|\Delta^j\| \leq 3\sum_{j\in J}\lambda_j\|\Delta^j\|$, that

$$
\begin{aligned}
\|\Delta_{J^c}\|_{2,1} &= \sum_{j\in J^c}\|\Delta^j\| \\
&\leq \sum_{j\in J^c}\frac{\lambda_j}{\lambda_{\min}}\|\Delta^j\| \\
&\leq \frac{3}{\lambda_{\min}}\sum_{j\in J}\lambda_j\|\Delta^j\| \\
&\leq \frac{3\lambda_{\max}}{\lambda_{\min}}\|\Delta_J\|_{2,1}.
\end{aligned}
$$

Combining these inequalities we find that

$$
\begin{aligned}
\frac{\Delta^\top\Psi\Delta}{\|\Delta_J\|^2} &\geq \frac{\Delta_J^\top\Psi\Delta_J}{\|\Delta_J\|^2} + \frac{2\Delta_{J^c}^\top\Psi\Delta_J}{\|\Delta_J\|^2} \\
&\geq \phi - \frac{2\lambda_{\min}\phi}{14\alpha\lambda_{\max}} - \frac{12\phi\|\Delta_J\|_{2,1}^2}{14\alpha s\|\Delta_J\|^2} \\
&\geq \left(1-\frac{1}{\alpha}\right)\phi.
\end{aligned}
$$

■

Lemma A.3. *Let $T \geq 16$ and $s \geq 16$. If ω and ω' are two elements of $\mathcal{N}'$ such that $\rho'(\omega,\omega') \geq \frac{Ts}{8}$, then the cardinality of the set $J(\omega,\omega') = \left\{j \leq s : \sum_{t=1}^T I\{\omega_{tj} \neq \omega'_{tj}\} > \frac{T}{16}\right\}$ is greater than or equal to $\frac{s}{16}$.*

Proof. Assume that $|J(\omega,\omega')| < s/16$. Then, denoting by $J(\omega,\omega')^c$ the complement of $J(\omega,\omega')$, and using that $|J(\omega,\omega')^c| \le s$, we get

$$\rho'(\omega,\omega') \le \sum_{j\in J(\omega,\omega')^c} \sum_{t=1}^{T} I\{\omega_{tj} \neq \omega'_{tj}\} + |J(\omega,\omega')|T < Ts/8,$$

which contradicts the premise of the lemma. ■

References

[1] D.A. Aaker, G.S. Day, and V. Kumar. *Marketing research*. John Wiley & Sons, 1995.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] F.R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[4] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[5] J.M. Borwein and A.S. Lewis. *Convex analysis and nonlinear optimization*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 3. Springer, New York, second edition, 2006. Theory and examples.

[6] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[7] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007.

[8] E. Candès and T. Tao. Rejoinder: "The Dantzig selector: statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), no. 6, 2313–2351; mr2382644]. *Ann. Statist.*, 35(6):2392–2404, 2007.

[9] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2002. Dedicated to the memory of Lucien Le Cam.

[10] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables Lasso. *Math. Methods Statist.*, 17(4):317–326, 2008.

[11] P.J. Diggle, P.J. Heagerty, K. Liang, and S.L. Zeger. *Analysis of longitudinal data*, volume 25 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, second edition, 2002.

[12] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[13] L. Dümbgen, S.A. van de Geer, M.C. Veraar, and J.A. Wellner. Nemirovski's inequalities revisited. *Amer. Math. Monthly*, 117(2):138–160, 2010.

[14] T. Evgeniou, M. Pontil, and O. Toubia. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science*, (26):805–818, 2007.

[15] C. Hsiao. *Analysis of Panel Data*. Cambridge University Press, 2003.

[16] J. Huang, J.L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. Manuscript. 2008.

[17] J. Huang and T. Zhang. The benefit of group sparsity. arXiv:0901.2962v2, 2009.

[18] V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems. Saint-Flour Lectures Notes, 2008.

[19] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Conference on Learning Theory, COLT-2008*, pages 229–238, 2008.

[20] P. J. Lenk, W. S. DeSarbo, P. E. Green, and M. R. Young. Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191, 1996.

[21] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

[22] K. Lounici, M. Pontil, A.B Tsybakov, and S.A. van de Geer. Taking advantage of sparsity in multi-task learning. In *Proc. of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

[23] A. Maurer. Bounds for linear multi-task learning. *J. Mach. Learn. Res.*, 7:117–139, 2006.

[24] L. Meier, S.A. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 2009.

[25] Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1):53–71, 2008.

[26] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.*, 2:605–633, 2008.

[27] A. Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000.

[28] G. Obozinski, M.J. Wainwright, and M.I. Jordan. Union support recovery in high-dimensional multivariate regression. *Annals of Statistics*, to appear.

[29] V.V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.

[30] G. Raskutti, M.J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. arXiv:0910.2042, October 2009.

[31] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam: Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 22, 2007.

[32] P. Rigollet and A. Tsybakov. Exponential Screening and optimal rates of sparse estimation. arXiv:1003.2654, March 2010.

[33] E. Rio. Moment inequalities for sums of dependent random variables under projective conditions. *J. Theoret. Probab.*, 22(1):146–163, 2009.

[34] V.K. Srivastava and D.E.A. Giles. *Seemingly unrelated regression equations models*, volume 80 of *Statistics: Textbooks and Monographs*. Marcel Dekker Inc., New York, 1987. Estimation and inference.

[35] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[36] S.A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[37] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[38] J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2002.

[39] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

[40] A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.*, 57:348–368, 1962.

[41] C.H. Zhang and J. Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

[42] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37(6A):3468–3497, 2009.